

分 类 号: TP391
研究生学号: 2013532046

单位代码: 10183
密 级: 公 开



吉 林 大 学

硕士学位论文

(学术学位)

基于情感倾向性的网络舆情分析及演化预测研究

Research on Analysis and Evolution Prediction of Network Public
Opinion based on Emotional Tendency

作者姓名: 孙培星

专 业: 计算机软件与理论

研究方向: Web 挖掘

指导教师: 彭涛 教授

培养单位: 计算机科学与技术学院

2016 年 5 月

基于情感倾向性的网络舆情分析及演化预测研究

Research on Analysis and Evolution Prediction of Network
Public Opinion based on Emotional Tendency

作者姓名：孙培星

专业名称：计算机软件与理论

指导教师：彭涛 教授

学位类别：工学硕士

答辩日期：2016 年 5 月 25 日

未经本论文作者的书面授权，依法收存和保管本论文书面版本、电子版本的任何单位和个人，均不得对本论文的全部或部分内容进行任何形式的复制、修改、发行、出租、改编等有碍作者著作权的商业性使用（但纯学术性使用不在此限）。否则，应承担侵权的法律责任。

吉林大学硕士学位论文原创性声明

本人郑重声明：所呈交的硕士学位论文，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：孙培生

日期：2016年5月25日

摘要

基于情感倾向性的网络舆情分析及演化预测研究

近年来,互联网社交工具的快速普及对我国社会产生了巨大的影响,并成为人们了解世界、交换意见的重要平台。在网络技术发展的同时,互联网和社会舆情逐渐融合,于是便产生了网络舆情的概念,网络舆情能够真实、快速的反映社会舆论,尤其是最近几年社交平台的迅速崛起,网络舆情在很大程度上引导着认识的基本认识,因此网络舆情的情感倾向及演化变的尤为重要,对政府的行政方式和决策机制都产生了极大的影响。所以,网络舆情分析技术便应运而生。

本文在传统文本分类算法支持向量机的基础上,针对网络社交媒介引入特定的情感词典,并将其应用在特征选择方面,构造文本倾向性分类器,使用该分类器判别微博的情感极性(正向或负向)。同时,进一步研究网络舆情的演化规律情况,即舆情热度的变化规律,综合考虑影响舆情热度的驱动因素,最后通过实验验证论文中采用的方法的可行性与有效性。本文的主要研究内容可以概括为以下四个方面:

1. 对 HowNet 中文词典重新整理与补充,尝试构建网络舆情分析的特定情感词典,为下文网络舆情情感分类器的构建奠定了一定的实验基础;
2. 对原始实验数据进行人工标注,并进行数据预处理;
3. 将情感词典应用到文本特征选择上,提出将词频法和互信息法相结合的特征提取方法,选取满足条件的特征并计算其权值,训练模型,并通过实验验证本文所提方法的有效性;
4. 利用训练得到的情感分类器对整体微博舆情进行极性判断,得到负向舆情信息集合,使用回归模型分析负面网络舆情的演化规律,并对网络舆情热度进行研究,找出影响舆情热度的因素,分析每种因素对舆情热度影响的显著性,建立多元线性回归预测模型,最后分析预测负向舆情与整体舆情热度的演化规律。

实验表明,在网络舆情情感分类方面,引入情感词典之后,所选取的特征更加具有领域性和代表性,再将词频和互信息方法相结合更能很好的表征数据,实验结果较单纯使用词频和互信息的特征选择方法更加有效。在网络舆情演化分析方面,把影响舆情热度的驱动因素作为多元线性回归模型的自变量,分析自变量的显著性以及它们之间是否存在多重共线性,并对模型的预测值和实际值做差值

分析，证明了模型应用于预测的可行性。最后，使用回归模型对负向舆情信息和整体舆情信息的热度做对比，分析了时序网络舆情的演化规律。

关键词：

倾向性分类，舆情分析，特征选择，多元线性回归

Abstract

Research on analysis and evolution prediction of network public opinion based on emotional tendency

The rapid popularization of Internet technology has a great impact on China's society, and becomes an important platform for people to understand the world and exchange the views. The Internet and social public opinion gradually fuse at the time of network technology development. And it causes the concept of network public opinion and network public opinion can be true, fast response to public opinion. As the rapid rise of the social platform in recent years, network public opinion largely cited by the basic understanding of the mass, so the emotional tendency and evolution of network public opinion becoming more and more important, which has a great impact on the government administration and decision-making mechanism. So, the network public opinion analysis technology has come into being.

Based on the traditional text classification algorithm-support vector machine algorithm (SVM), this paper introduces the specified emotion dictionary and applies it on feature selection for the network of social media. Then this paper structures text tendency classifier and uses the classifier to discriminant blog sentiment polarity (positive or negative). At the same time, we further study of the evolution of network public opinion - heat of public opinion changes – and consider the influence factors of the heat of public opinion. Finally, we verify the feasibility and effectiveness of the method proposed in this paper through experiment. The main content of this paper is described as the following:

1. Rearranging and supplementing dictionary based on HowNet Chinese dictionary, and trying to build specified emotion dictionary for the network public opinion analysis, which provided the experimental basis for the construction of the network public opinion text classifier.

2. Manually labeling of the original experimental data, and doing the work of data preprocessing.

3. According to the characteristics of network public opinion in this paper, we

applied the sentiment dictionary to text feature selection, and put forward the method of combining frequency method with mutual information - the linear combination of feature extraction method, then selected the feature which meet the conditions and calculated the weight of it, trained the model, and verified the effectiveness of experiments proposed in this paper.

4. Using the emotional classifier trained to judge polarity on the whole micro-blog, obtaining the set of negative micro-blog, then using regression model to analyze the evolution of negative network public opinion, and find the influence factor of public opinion heat. Next, analyzing the significance of each factor on the heat of public opinion, establishing multiple linear regression prediction model, and finally analyzing and forecasting the evolution of the negative public opinion and the overall public opinion heat.

Experimental results show that the features we selected is more field, representative and combining the word frequency and mutual information method can well characterize the data in the Internet public opinion emotion classification after introducing of sentiment lexicon. The result is better than only use word frequency and mutual information method in feature selection. In the evolution analysis of Internet public opinion, this paper take the driving factors influencing the heat of public opinion as independent variable of multiple linear regression models, analyzing the significant argument of independent variables and whether there is multi-collinearity. Then the model is tested by residual error, and the model is available for prediction. Finally, the paper compares the heat of the negative information with the heat of the overall public opinion information using regression model, and analyzes the evolution law of the time sequence network public opinion.

Keywords:

Orientation Classification, Public Opinion, Feature Selection, Multiple Linear Regression

第 1 章 绪论	1
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	3
1.3 本文研究工作.....	6
1.4 本文组织结构.....	7
第 2 章 相关理论及关键技术	8
2.1 网络舆情分析介绍及相关理论	8
2.1.1 网络舆情技术.....	8
2.1.2 网络舆情的传播.....	9
2.1.3 有关网络舆情的其他研究	12
2.2 倾向性分析.....	12
2.2.1 研究分类.....	12
2.2.2 技术分类.....	13
2.3 多元线性回归分析.....	16
第 3 章 基于情感词典的网络舆情倾向性分类研究	18
3.1 基于 HowNet 的情感词典构造	18
3.1.1 HowNet 简介.....	18
3.1.2 基于 HowNet 的词语相似度计算	19
3.1.3 知网 HowNet 词典的扩展	20
3.2 基于情感词典框架下的网络舆情倾向性分类	23
3.2.1 数据预处理.....	24

3.2.2 文本特征提取.....	26
3.2.3 线性组合特征选取算法	27
3.2.4 特征权重计算.....	31
3.3 实验测试及结果分析.....	33
3.3.1 实验环境及参数选取.....	33
3.3.2 分类器学习方法的选择	33
3.3.3 评估标准.....	34
3.3.4 算法测试及对比分析.....	35
第4章 时序信息的网络舆情演化规律模型	39
4.1 网络舆情演化.....	39
4.2 网络舆情演化的驱动因素	40
4.2.1 网络舆情驱动因素介绍	40
4.2.2 网络舆情驱动因素分析	41
4.3 多元线性回归预测.....	43
4.4 实验结果及分析.....	44
4.4.1 问题描述.....	44
4.4.2 实验过程.....	44
第5章 总结与展望	50
参考文献	51
作者简介及在学期间科研成果	55
致 谢	56

第1章 绪论

1.1 研究背景和意义

近年来,随着大数据、云计算等计算机技术的快速发展,互联网在世界范围内得到了广泛的普及。2015年7月,CNNIC发布的第36次《中国互联网络发展状况统计报告》报告^[1],显示:到2015年6月为止,我国已经达到6.68亿的网络规模,达到48.8%的互联网普及率,半年期间一共增加1894万人的网民。互联网对个人生活方式的影响进一步深化,人们通过各种网络社交媒体在互联网中传播信息,其中新闻、博客、论坛、微博、微信等互联网应用在信息传播中占据着重要的角色,每天源源不断的产生大量的信息数据,这种信息传播方式使得网络舆情有了快速的发展。

网络舆情是在一定社会空间中,网民以网络平台为中介对公共问题、公共事件、社会问题所发出的观点、态度、信念和价值观。传统的舆情主要指民间的大众言论,常见于日常的街头巷尾,这种舆情需要街头暗访,民间调查,难以获取,而且成本较高、效率低下。如今,随着互联网技术的发展,尤其是移动互联网技术的快速普及,人们纷纷在各种社交媒体、平台上发表自己的看法,表达自己的态度,言论更加自由,而且信息更加丰富。所以,可以使用网络技术抓取网络舆情,节省了大量的人力物力,同时提高了抓取效率。

网络言论在狭义上讲包括新闻媒体的言论和网民言论。其中前者是传统媒体(如:报纸、杂志、期刊等)在互联网中的延伸,它一般会受到相关部门的监督和管理。所以,这种媒体言论严格的讲不能完全代表舆情。网络言论主要是在微博、论坛、新闻评论等具有交互性质的网络平台进行传播,包括信息的转发、点赞、浏览、留言、回帖等行为,可以快速的集中反映公众的意见,具有自发性与自主性,不受其他因素的约束,更能真实的反应舆情的发展。网络舆情是由网络社交平台中的言论发展而来的,但不是说所有的网络言论都会发展演变成为网络舆情,这在很大程度上取决于该言论的内容是否具有敏感性以及该言论在网络平台的活跃程度,与此同时,还取决于网络言论是否对政府、社会和人民构成危害。同时,网络的开放性和和隐匿性为舆情提供了公共的场所,如果我们能够提前将

危害社会的不良舆论进行有效的抑制,就可以发挥前瞻性和预见性,能够很好的把握网络舆情的主动权。常见的舆情如:流感的传播、雾霾天气、社会道德问题、军事活动等敏感和热点问题。习近平主席曾经强调,我国要建设成网络强国,积极做好网上舆情工作。只有及时把握网络的这些舆情信息,政府才能够在最短的时间内了解大众民意,并根据局势快速的做出正确的决策,维护好社会的稳定。

在海量的网络数据当中,通过探测并发现网络舆情中的活跃主题,找到舆情发展的思路,从海量信息中找到目标信息,以便更好的应对舆情分析。同时,对网络舆情进行高效的监督管理和维护,对突发的舆情能够迅速做出决策,化解网络舆情危机,保持国家政府社会稳定发展,为构建和谐的网络环境具有重要的现实意义。而且,随着大数据时代的到来,网络舆情数据逐渐呈现出大数据的特征。通过观察,我们看到,互联网的自由性与开放性使得社会中的各种群体能够方便的发表自己的观点,表达自己的态度,使得网络数据量急剧增多。而且,伴随着社交方式的多样性,网络舆情可以不仅仅局限于文本表达言论,同时还有图片、视频、音频等数据格式。数据量的增大,使得舆论多元化,网络舆情变化多变。由于网络信息量巨大,所以仅仅依靠人工的办法很难获取,现阶段工业界和学术界都在舆情分析领域做了很多的实现与研究,如国内的有邦富^[2]、红麦软件^[3]等公司,国外的有 Buzzlogic^[4]、Nielsen^[5]、Reputation Defender 等。所以,近年来一些学者们开始研究网络舆情的特点,希望能从中找到新的切入点进行深入研究。在网络舆情分析中,文本倾向性分析是一个研究较为深入且广泛的方向。

网络中的舆情大部分都是针对某一主题进行展开讨论的,这种舆情称作主题网络舆情。主题网络舆情一般在态度上具有一定的情感倾向,可以通过分析判断主题舆情的情感倾向性,来判断公众对该主题活动的观点和态度,如果发现危害社会和国家等不良因素,相关部门能够及时科学的制定方案,处理相应的网络舆情,使网络舆情向健康积极向上的方向发展^[6]。舆情分析中的文本倾向性研究主要是指文本的分类研究,它在各个领域有着广泛的应用,如购物网站中的商品评价、博客、新闻评论、论坛等领域。基于文本倾向性分析的文本挖掘,需要对文本的语义进行一定的分析理解,并在此基础上挖掘作者的意见和态度等知识。在这种海量舆情环境中,文本倾向性分析(sentiment classification)定会有用武之地,挖掘人们的评论文本,自动检测网络中的评论是正面的还是负面的。文本倾

向性分析作为网络舆情的一个重要技术,将被用于微博等社交平台上的舆情分析。总之,文本倾向性研究具有重要的社会价值,它可以与其他学科结合并渗透到其中为各个学科的发展提供了发展动力,尤其是在大数据环境下,如果文本倾向性分析能够在网络舆情领域得到合适的应用,必然具有很大的潜在价值。

1.2 国内外研究现状

关于舆情分析方面的研究,我国在这方面的研究发展较慢。在 2000 年,我国才有一些学者开始相关领域的研究,在 2003 年之后才步入舆情分析的正轨,在 2008 年舆情分析技术逐渐走向成熟。同时,舆情分析在各个国家也是重要的研究课题。下面从文本倾向性、舆情分析系统和网络舆情演化三个方面进行阐述。

在文本倾向性方面:文本倾向性分析^[7,8]属于数据挖掘和机器学习领域中的关键算法^[9],是自然语言处理(NLP)领域的研究热点之一,近年来随着人们对网络文本倾向性分析工作的不断深入,文本倾向性的研究范围不断扩展,并逐渐在学术界占有重要的地位,很多国内外学者很早就有关于这方面的研究。国内一些学者在微博用户特征、微博话题、web 评论信息、语义事件等方面进行了有关文本倾向性的研究。姚天昉等人^[10]从观点的角度出发,对意见挖掘进行了定义,然后阐述了意见挖掘的目的,针对文本意见从不同角度进行了综述。清华大学的周立柱等人^[11]站在技术的角度,对用户发出的评论进行分析,挖掘出文本倾向的演变规律,进而更好的理解用户的消费习惯,确定舆情发展趋向。Yue Lu 等人^[12]在主题倾向性混合统计模型的基础上,采用半监督学习方法计算得出文章的整体情感倾向性。2008 年的信息检索专业委员会推出了中文倾向性分析评测,这大大的推动了我国情感倾向性分析的研究,推进了倾向性分析的发展研究,同时还产生了大量的优秀学术论文。针对微博等短文本特征稀疏问题研究,Zelikovitz^[13]在计算文本相似度时利用没有标注的背景知识进行改进。王永恒等人^[14]引入本体库的概念解决短文本的分类问题。文献^[15]提出直推式的机器学习新颖方法。通常认为,传统的文本分类是基于机器学习的文本倾向性的研究方法的重要部分,基于机器学习的文本倾向性分析方法首先是通过标注一些文本的倾向性(正向或负向),生产满足一定格式的文本数据,再将这些标记的文本作为训练数据,通过使用机器学习、数据挖掘的分类方法构造一个情感分类器,最后使用该分类器对

测试数据进行分类,计算出测试数据的情感倾向性及其情感强度。早在2002年,Pang等人^[16]就使用朴素贝叶斯算法(Native Bayes Algorithm)、支持向量机算法(Support Vector Machines Algorithm)、和最大熵算法(Maximum Entropy Algorithm)对文本倾向性进行了实验验证和对比实验,最后得出的结果是这三种算法的效果相差不大,相对而言,支持向量机算法的效果较其它两种算法较好一些。Mullen等人^[17]使用支持向量机方法,建立文本分类模型,对文本进行倾向性分析。除了将分类理论应用在倾向性分析以外,一些学者还将语义模式进行文本倾向性分析,比如Yi^[18]提出使用情感分类器(Sentiment Analyzer),主要使用句法分析器实现对实验数据文本句子的语法分析,接着利用情感词典(Sentiment Dictionary)和情感模式库两种方法的结合,分析上述句子的语义关系并计算得到该句话的文本倾向性。此外,还有一些学者提出,对文本中情感词汇的语义倾向进行度量,求其所有词汇的语义倾向性值的平均值,最后得到整体文本的倾向性度量结果。Wilson等人^[19]通过分析否定词、副词、语境等相关因素,构建一个情感词典,并以此为标准判定情感倾向性。

在网络舆情分析系统方面:网络舆情倾向性分析的主要研究对象是web文本,通过分析web文本内容所表现出的倾向和态度,然后通过某种模型或情感度计算方法对活动或事件的看法和态度进行量化,从而得出该看法或评论对该事件的情感倾向性。文本倾向性大致包括三个部分:一、人工标识出文本中能够体现正负情感的词汇;二、计算出标识的情感词的情感倾向性和强度;三、对整个文本的情感倾向性进行计算,并作出情感极性的判断。

通过对国内外有关舆情分析的研究发现,Web舆情分析主要借助数据挖掘和机器学习中的各种技术和算法思想对信息进行搜集,预处理,发现舆情,分析舆情,生成分析结果报告等环节。Buzzlogic是一家借助数据分析技术的网络舆情分析的国外公司,其提供的“BuzzLogic Insights”服务能够对博客等平台进行高时效的、全方位的、多角度的实时监测,同时对监测数据实时分析给出舆情动态。这些分析结果能够帮助企业了解用户需求,分析用户行为习惯,发现市场行情提供重要的依据。

现阶段,国内对Web舆情的研究还不够深入,舆情分析技术还不是太成熟。其中一个原因与中文的语言特点有关,中文相对英文在分词方面相对复杂和困难,

而且中文含义较多,在不同的语境有不同的含义,语义复杂多变,所以说让计算机像人一样理解中文是十分困难的。邦富是国内做的较好的公司之一,它是一家长期从事舆情分析、搜索研发的企业,也是国内最大最专业的网络舆情系统开发商。该公司的互联网舆情监控系统包括:分钟级的实时采集技术、多级分类机制、信息自动去重功能、独有网页正文提取技术、精确的智能摘要、信息趋势分析、基于 Web 的系统管理平台等功能。其中,有多项技术已经转化为科技研究成果。但是,这种舆情检测系统通常擅长于抓取网页,在社交网络(博客、微博、BBS等)中,效果十分明显,主要是因为网络社区中的舆情主要靠人工分析而来。

在网络舆情演化方面:当前网络舆情的演化规律相关研究主要侧重于宏观的层面,即基于经验的研究,而缺乏对微观的研究。国内学者刘怡君等人^[20]针对舆情的发展阶段概括了三个阶段,分别为潜伏期、活跃期和衰退期。在舆情热度方面,刘勘等人^[21]使用隐马尔科夫链建立模型预测舆情的热度演化。一些学者使用时间来分析舆情的变化规律,建立模型进行分析^[22,23,24]。Tam^[25]等人根据不同类型的网络提出了不同类型的网络模型。Shen B^[26]等人考虑到初始状态的意见和主题的接触过程,提出一个观点形成模型。总的来说,国内有关舆情演化的研究主要从经验角度集中于对舆情演化阶段及演化流程的总结,这种研究方法缺乏数据的支持,不能完全反应网络舆情的整个发展过程。而国外学者更加注重人们的观点演化规律,对每个个体建立演化模型,这种研究方法具有很好的参考意义。

随着大数据时代的到来,社交网络有了快速的发展,尤其是移动互联网的发展让人们轻而易举的在社交网络上浏览和评论。为了研究网络舆情的变化趋势,我们可以从网络中的数据开始抓起。李国杰院士曾经提到,“数据的背后是网络,网络的背后是人”,研究网络数据实质上在研究网络中人与人的社会关系。以微博为例,微博中的每一个用户看做一个节点,用户之间可以进行互相关注、评论、转发、点赞等操作,如果用户 A 关注了用户 B,则可以画一条由用户 A 指向用户 B 的有向边,如果将微博网络中的所有用户之间的这种关系用有向边连接起来,就会形成一个巨大的社会网络图。可以想象,在这个巨大的图中,舆情信息会在节点中频繁的流动,最终构成一个舆情网络。

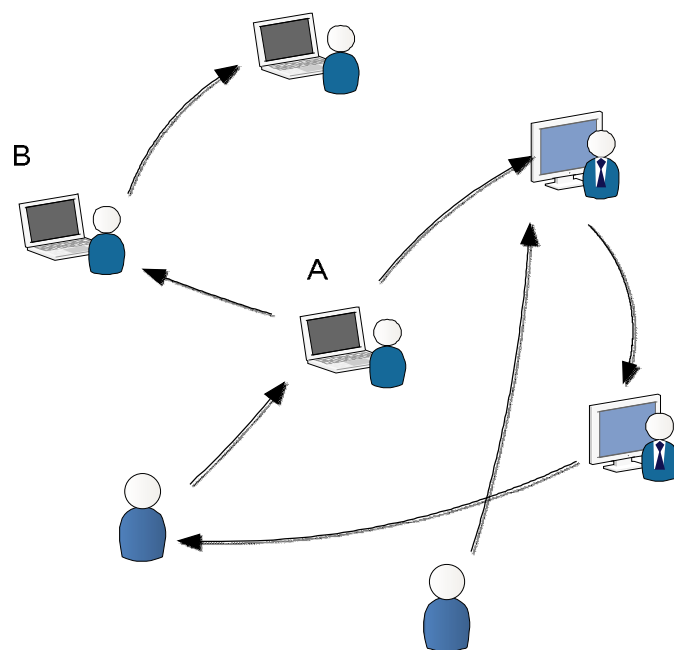


图 1.1 微博用户关注关系数据流图

因此,对于这样的网络必然存在很大的研究价值。面对互联网中的海量数据,有效的分析、发现、挖掘网络舆情是当前学术界和工业界的主要工作。因此,舆情分析将是未来的主要研究方向。

1.3 本文研究工作

本文首先对网络舆情数据进行情感倾向性分析,然后再对网络舆情演化的影响因素进行分析。前者通过人工方法对预处理的数据进行标注,构建领域内的情感词典,将该情感词典与特征选择方法相结合选择合适的特征,然后使用支持向量机算法(SVM)对处理的网络舆情数据进行训练,得到最终的网络舆情分类器,并用该分类器对待分类数据进行分类,得到分类的准确率、召回率和F值。后者在前者的基础之上,借助已训练好的舆情分类器对所有舆情数据进行极性判断,得到负向舆情数据。分析舆情的演化影响因素,构建多元线性回归模型,并对模型进行参数检验,找出影响舆情热度发展趋势的关键因素,最后分析整体舆情和负向舆情热度的演化规律,为政府对舆情的发展演化检测提供重要的决策参考。

1.4 本文组织结构

本文具体安排如下：

第1章，绪论。主要介绍了舆情分析的研究背景和意义，同时针对舆情分析的国内外发展现状进行简单阐述，并简要说明了本文的主要研究工作和组织结构。

第2章，相关理论和关键技术。这一章主要介绍网络舆情分析中的相关概念和基本理论。

第3章，基于情感词典的网络舆情倾向性分类研究。这是本论文的重点内容，首先介绍了情感词典的构建，特征的选择方法，提出词频和互信息结合的线性组合特征选择方法，然后进行模型训练及预测。

第4章，时序信息的网络舆情演化规律模型。本章使用上一章的情感分类器得到负面舆情信息，分析了影响舆情热度的多个因素，使用多元线性回归模型对网络舆情热度进行演化预测分析。

第5章，总结与展望。对全文做总结，并对未来的研究方向做出讨论，明确今后的发展目标。

第2章 相关理论及关键技术

2.1 网络舆情分析介绍及相关理论

2.1.1 网络舆情技术

网络舆情是指在互联网中人们对社会问题的不同看法的网络舆论，它是以网络为载体，以事件为核心，在网上传播观点、态度、意见的一种信息扩散方式。网络舆情分析理论是一个复合学科，其研究涉及到新闻传播学、社会学、情报学、计算机科学等多个领域^[27]。网络舆情分析技术主要包括内容分析法、网络舆情信息采集与提取、网络文本倾向性分析技术、网络舆情话题发现与追踪技术、文本数据挖掘法。

1. 内容分析法。这是一种有效的社会科学研究方法，主要在三个方面得到应用：一是网络中的舆情信息；二是对网络舆情信息的态度和情绪进行推断；三是分析网络舆情信息并得出趋势变化。

2. 网络舆情信息采集与提取。网络舆情主要通过网站媒体（博客、微博、论坛、新闻等）进行传播，这些网站大多数使用动态网页，并以数据库为基础，当需要注册、登陆、管理等加载功能时，需要调用数据库进行动态抽取。国内学者梅雪等人^[28]提出了一种称为“Wrapper”的全自动的抽取方法，在一定程度上提高了处理的准确率。

3. 网络文本倾向性分析技术。网络中新闻评论、微博评论等的文本通常具有一定的情感倾向。对网络舆情分析实质上就是对网络中的文本分析其所具有的情感倾向，所以文本倾向性分析是网络舆情分析的重要技术之一。

4. 网络舆情的主题发现与跟踪^[29]。如何在海量信息中找到舆情热点信息和敏感话题是网络舆情分析的主要研究方向。

5. 文本数据挖掘法。该方法借助数据挖掘中的关联规则、文本聚类、文本分类等技术和算法，从网络数据中挖掘出可用的知识。

2.1.2 网络舆情的传播

通常，舆情在网络中的传播形式类似于传染病在人群中的传播，可以做出这样的比喻，将网络中的负面消息看做是传染疾病，则可以参照已有的传染病传播模型建立舆情的复杂网络模型。经典的传染病模型^[30]有 SIS、SIR 和 SIRS 模型。假设条件是：人口总数不变，为 N ；正常个体与感染者接触后会被感染，而且他们之间有一个固定的感染概率 λ ，设定一个阈值为 λ_c 。如果实际传染速率大于 λ_c 时，传染病即将爆发；同时假设每个个体具有相同的恢复率 γ 。

1. SIS 模型

假设 SIS 模型在人群中只存在两种状态：易感状态 S (Susceptible) 和传染状态 I (Infected)。假设某个个体当前属于健康状态，当该个体接触到感染者时，会以感染率 λ 的概率被感染，并从易感状态转移到传染状态，此时该个体可以传染给接触到它的个体，并以恢复率 γ 恢复到易感状态。具体的传染规律如下图 2.1 所示：

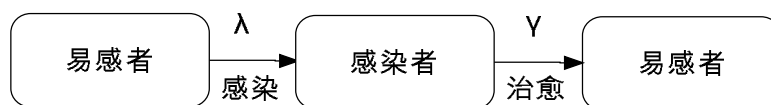


图 2.1 SIS 模型演化规律图

像流感等传染疾病的传播一般都符合 SIS 模型的规律。现在定义 $S(t)$ 、 $I(t)$ 分别是易感者和感染者在人群中的比例，模型的具体传播微分动力学方程：

$$\left\{ \begin{array}{l} \frac{dS(t)}{dt} = -\lambda I(t)S(t) + \gamma I(t) \\ \frac{dI(t)}{dt} = \lambda I(t)S(t) - \gamma I(t) \\ S(t) + I(t) = 1 \end{array} \right. \dots\dots\dots (2.1)$$

2. SIR 模型

上述中的 SIS 模型描述了大部分易感者和感染者之间的传播规律，但是对于一些疾病（如天花等）易感者在获得感染者的感染之后会具有了免疫力而不会再次受到感染。因此，Reed 等人在 1920 年提出了一种新的传播模型--SIR 模型。该模型是在易感者和感染者的基础上添加了免疫者 R (Recovered)，当感染者恢复为健康的状态时，再次接触感染者将不会受到感染。SIR 模型的具体演化图 2.2

如下:

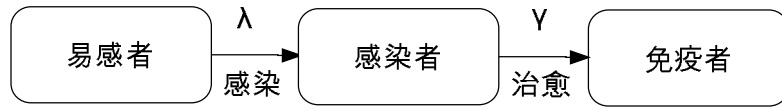


图 2.2 SIR 模型演化规律图

同样地, 分别用 $S(t)$ 、 $I(t)$ 和 $R(t)$ 定义为易感者、传染者和免疫者在人群中的比例。用平均场方程表示模型的传播微分动力学方程:

$$\left\{ \begin{array}{l} \frac{dS(t)}{dt} = -\lambda I(t)S(t) \\ \frac{dI(t)}{dt} = \lambda I(t)S(t) - gI(t) \dots\dots\dots(2.2) \\ \frac{dR(t)}{dt} = gI(t) \\ S(t) + I(t) + R(t) = 1 \end{array} \right.$$

3. SIRS 模型

上述描述的 SIR 模型还不能完全解释一些疾病, 比如结核病, 患者治愈后具有了免疫能力, 但是这种免疫能力只能持续几年或几十年, 在某个时期这种免疫能力会消失, 恢复后的个体可能再次受到感染。因此, 一些学者又提出了 SIRS 模型。假定具有免疫能力的个体以 δ 的概率丧失免疫能力。下图是 SIRS 模型的演化过程。

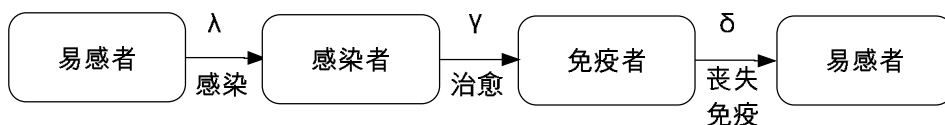


图 2.3 SIRS 模型演化规律图

接着借助平均场方程来表示模型的传播微分动力学方程:

$$\left\{ \begin{array}{l} \frac{dS(t)}{dt} = -I(t)S(t) + dR(t) \\ \frac{dI(t)}{dt} = I(t)S(t) - gI(t) \dots\dots\dots(2.3) \\ \frac{dR(t)}{dt} = gI(t) - dR(t) \\ S(t) + I(t) + R(t) = 1 \end{array} \right.$$

借鉴疾病传播模型，可以构建网络舆情的 SIS 模型和 SIR 模型，这两种模型都将传播网络拓扑假定为规则网络，只是传播的规则不同。下面介绍 S 、 I 、 R 的具体含义：

- Ⅰ S ：未上传负面信息的网络媒介（正常的节点）
- Ⅰ I ：已经上传负面信息的网络媒介（感染的节点）
- Ⅰ R ：已经上传负面信息但失去感染能力的网络媒介（具有免疫能力的节点）

下面是网络舆情信息的传播过程。首先，随机选择一个节点作为上传负面信息的网站（即感染的节点），其他节点为健康节点，当一个网站的浏览者浏览到该负面信息时（例如访问新浪微博中的马航失事事件），该浏览者可能会基于习惯或兴趣将该负面信息转发到其他网站（节点），从而使该网站以概率 λ 变成感染的节点；同时，由于每个网站都有网络管理员，会定期清理网上的部分负面消息，所以感染的网站会以概率 γ 恢复，变成健康网站（节点）。也就是说，如果一个网站的感染率越大，恢复率越小，那么这个网站的负面消息就越容易传播到更多的网站中。以上信息的传播过程就是网络舆情的大致演变流程。下图 2.4 是具体的信息传播过程。

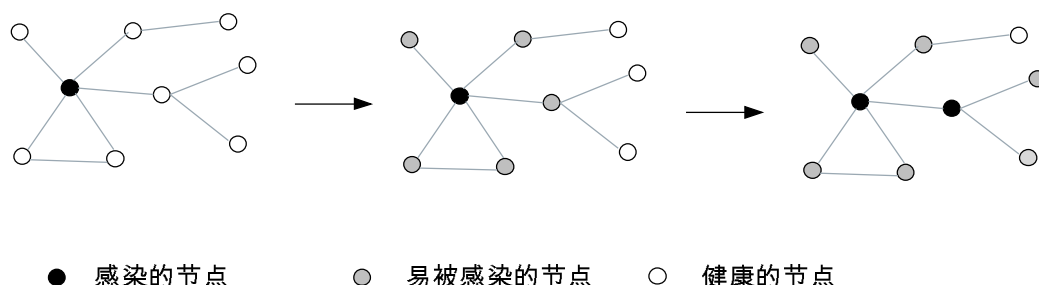


图 2.4 负面舆情传播过程

不少学者认为，舆情在网络中的传播模式与传染病在人群中的传播模式类似。

近年来,随着微博等社交媒体的迅速发展,有关微博数据的研究也受到更多学者的关注,一些学者认为微博中的“关注”是网民互动的重要方式,其中“关注者”与“被关注着”是一种“传播”和“接受”的关系。因此,把微博用户看做是节点,把关注看做是边,那么这个微博可以看做是一个复杂的网络结构。

2.1.3 有关网络舆情的其他研究

网络舆情借助网络新媒体在传统的社会舆情基础上发展起来,同时,结合已有的文本处理算法和自然语言处理方法,因此,网络舆情具有很大的研究价值。下面就几个方面简单介绍:

1. 网络舆情与话题相关性的判定。在获取舆情信息时,为了得到相关度较高的网页内容,要综合考虑网页内容和 URL 链接^[31,32]的价值信息,即从主题的相关度和链接的权威性进行考虑。

2. 网络舆情的聚类 and 分类^[33]。以舆情信息的语义相似度为核心,进行网络舆情信息的聚类 and 分类,进而对舆情热点进行监测和跟踪。

3. 基于情感倾向性分析(主要指文本)的网络舆情。实质上是对网页的文本进行情感分类,确定文本的情感极性,进而确定整个文本的情感倾向。具体介绍见下文。

2.2 倾向性分析

倾向性分析又叫情感分析(Sentiment Analysis, SA)^[34,35],主要应用在新闻资讯、微博等自然语言处理领域中,主要是挖掘文本对象中所蕴含的情感倾向。情感倾向按照褒贬一般包括正向、中立、负向。正向表示积极向上的态度,如:“高兴”、“喝彩”、“希望”等让人感到愉悦的词语;中立表示一种客观的态度,没有褒贬强烈上的情感倾向,不偏不倚没有包含任何情感色彩,如:“按部就班”、“无声无息”、“走马上任”等中性词语;负向表示让人郁闷、生气的词语,如:“焦虑”、“失望”、“悔恨”等。

2.2.1 研究分类

由于倾向性分析的研究对象是中文文本,中文有自己的语法结构,所以它涉及到语言分析层面,包括:词语层级、句法层级和篇章层级。

1. 词语层级。词语级的倾向性分析主要针对情感表构建及扩充，但是随着研究的不断深入，发现情感词具有强烈的领域依赖性。也就是说，通用的情感词典很难存在，需要针对特定的领域构建特定的情感词典。

2. 语句层级。从粒度的角度讲，语句层级倾向性分析相对于词语层级较粗。国外学者 Kim 和 Hui^[36]通过识别并计算每个句子中的情感词，然后对每个句子的情感极性进行整合得出语句的综合情感极性。大部分关于情感词的研究都是假定每个句子中只有一种情感倾向的情感词，但是在实际应用中，一个句子通常包含不止一种情感倾向的情感词。面对这样的问题，国内外一些学者已经开展了相关的研究。这样复杂的情感词倾向性仍是当前需要急需解决的问题。

3. 篇章层级。篇章层级是一个比词语级和语句级粒度更大的一种情感倾向性分析，其倾向性更加具有宏观性，得出篇章的倾向性也是文本倾向性极性判断的最终目的。对篇章级的文本分析不仅依赖于词法和句法分析技术，还需要跨句的指代消解。

2.2.2 技术分类

本文主要介绍两种网络舆情情感倾向性分类的方法：基于机器学习算法和基于情感词典的方法。下面是这两种方法的具体介绍。

1. 基于机器学习算法

网络舆情的倾向性分析实质是文本的倾向性分析，所以，我们可以将网络舆情的倾向性分类看做是机器学习中的二元分类问题（正向和负向），常用的文本分类算法有：支持向量机（SVM）^[37]、朴素贝叶斯^[38,39,40]、最大熵^[41]等算法。虽然它们具体实现的细节不同，但是它们的整体架构和流程是统一的。如下图所示：

（一）支持向量机

支持向量机（SVM）是数据挖掘中的分类算法，它的思想主要来自最优分类面。最优分类面要求分类面不仅能将两类数据正确的分开，而且还要确保分开的两个类别间隔最大。最大间隔法是构造最优分类面的方法之一，该方法通过求两类数据的支持向量，并求出穿过支持向量且能够区分两类数据的两条直线，同时保证这两条直线间隔最大化，即将分类问题转化为对偶问题，计算更加简单。为了提高算法的泛化能力，在非线性分类中引入惩罚因子，允许一定的分类错误，最终得到可伸缩的间隔，该方法的技巧是把一个复杂的最优化问题转化为内积运

算的形式。下图 2.5 是线性可分的例子，图中“+”和“-”分别代表样本数据的两类数据， H 是最理想的分界面， H_1 和 H_2 代表能够区分两类数据并且距离两类数据最近的分界面。由于支持向量机模型要保证经验风险最小，所以要确保两类数据的间隔 M 最大，所以在众多的分界面中，选择 H 作为最好的结果。

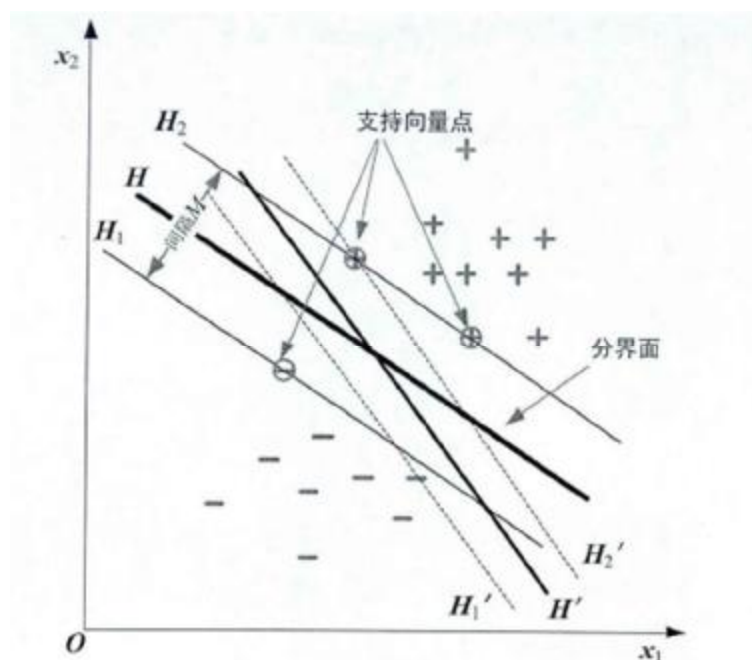


图 2.5 支持向量机分类算法的原理图

假设线性近似可分的样本集是 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, $x_i \in R^f$ (数据集为 f 维), 类别标识 $y_i \in \{1, -1\}$, $i \in [1, n]$ 。则该 f 维数据的线性判别函数的一般形式为

$$f(x) = w^T x + b \dots\dots\dots(2.4)$$

其中 w 为 f 维向量, b 为常量。样本的学习问题转化为约束最小化问题(Constrained Minimization Problem),

$$\min \frac{\langle w \cdot w \rangle}{2} + C \sum_{i=1}^n \xi_i \dots\dots\dots(2.5)$$

$$\text{s.t. } y_i (< w \cdot x_i > + b) \geq 1 - \xi_i, \quad i = 1, 2, 3, \dots, n$$

$$\xi_i \geq 0$$

因为目标函数是二次的 (quadratic) 和凸的 (convex), 所以可以使用标准的拉格朗日乘子 (Lagrangian multiplier) 来解决。

$$L(w, b, a) = \frac{1}{2} \langle w \cdot w \rangle + C \sum_{i=1}^n x_i - \sum_{i=1}^n a_i y_i ((\langle w \cdot x_i \rangle + b) - 1) \dots\dots\dots(2.6)$$

其中 $a = (a_1, a_2, \dots, a_n)^T$ 为拉格朗日乘子，然后分别求拉格朗日函数对 w, b, ξ 的偏导数，根据 Kuhn-Tucker 条件可得：

$$\nabla_w L(w, b, a) = 0 \dots\dots\dots(2.7)$$

$$\nabla_b L(w, b, a) = 0 \dots\dots\dots(2.8)$$

$$\nabla_x L(w, b, a) = 0 \dots\dots\dots(2.9)$$

得到：

$$\sum_{i=1}^n y_i a_i = 0, w = \sum_{i=1}^n y_i a_i x_i, C = a_i + b_i \dots\dots\dots(2.10)$$

将公式(2.10)带入转化为对偶形式：

$$\max \sum_{j=1}^n a_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j a_i a_j (x_i \cdot x_j) \dots\dots\dots(2.11)$$

$$\text{s.t. } \sum_{i=1}^n y_i a_i = 0, 0 \leq a_i \leq C, i = 1, 2, \dots, n$$

其中， x 是训练样本， y 是类别标识， a 是拉格朗日乘子， C 是惩罚因子。

（二）朴素贝叶斯

朴素贝叶斯（NB）分类源自贝叶斯理论，是一种基于概率的学习方法，这是一种十分简单的分类算法。利用类别的先验概率来计算未知文本或词语所属类别的概率，这种算法虽然简单但是一般都能取得不错的效果。同时，该算法有一个假设条件：属性之间相互独立，该条件在实际应用当中通常是不满足的。这一特点使得朴素贝叶斯在应用上有一定的局限性。

（三）最大熵

最大熵原理是指，当对一个未知事件的预测时，且该事件是一个随机事件，我们带有主观色彩的进行假设，而应满足它的所有已知条件。只有满足这样的条件才能让概率分布更加均匀，预期的风险才能更小。此时的概率分布的信息熵最大，“最大熵”的概念由此而来。最大熵模型的优点是只需选择较好的特征，而不需要考虑如何使用这些特征，特征选择灵活，方便使用，模型的可移植性强。

缺点是时间和空间开销较大，数据稀疏问题严重。

2. 基于情感词典

对于中文的舆情分析，其情感词典主要是基于 HowNet，HowNet 是中文文本处理方面比较权威的资源，是一个通用的情感词典。但是对于网络文本来讲，通用的词典往往不能完全满足实际需要，还需要添加网络上出现的新的情感词（包括表情符号等）。

2.3 多元线性回归分析

(1) 多元线性回归分析

多元线性回归^[42]是用来确定因变量 Y 与自变量 X_1, X_2, \dots, X_n ($n \geq 2$) 之间的相互依赖的统计分析方法。其方程表示如下：

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n + e \dots\dots\dots(2.12)$$

其中， Y 为因变量， X_1, X_2, \dots, X_n ($n \geq 2$) 为自变量， $\beta_1, \beta_2, \dots, \beta_n$ ($n \geq 2$) 为对应 X 的回归系数， β_0 为常量项， $\varepsilon \sim N(0, \sigma^2)$ 。采用最小二乘法^[43,44]求回归系数。为获取样本，对 Y 和 X_1, X_2, \dots, X_m 分别进行 n 次独立观察，得到

$$(Y_i, X_{i1}, X_{i2}, \dots, X_{im}), i = 1, 2, \dots, n$$

那么多元线性回归分析的矩阵形式为：

$$Y = X\beta + \varepsilon \dots\dots\dots(2.13)$$

$$\text{其中, } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

$$\text{接下来求估计值, 设}\beta\text{的估计值}\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{bmatrix}, Y\text{的估计值}\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \text{使用最小二}$$

乘法得到下面公式的多元线性回归模型方程：

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_mx_m \dots\dots\dots(2.14)$$

(2) 多元线性回归模型的检验

模型检验是回归模型的一个重要环节，由上一节公式（2.14）可以得到多元线性回归方程，本节将介绍回归方程的检验方法。通常情况，使用 R 检验、T 检

验和 F 检验分别对模型进行拟合优度检验、回归系数显著性检验和回归方程显著性检验。

1. R 检验

R 检验能够反映回归方程对样本数据的综合拟合效果， $R^2 (0 \leq R^2 \leq 1)$ 是度量拟合度的统计量，如果 R^2 的值接近于1，说明回归方程对样本数据拟合的很好，反之，拟合度较差。

2. T 检验

T 检验主要判断自变量是否是因变量的一个显著性影响因素，也检验每个回归系数是否有意义。其构造统计量

$$t = \frac{\hat{b}_i - b_i}{S_{\hat{b}_i}} \dots\dots\dots(2.15)$$

其中，t 为统计量， $\hat{\beta}_i$ 为回归系数估计值， β_i 为回归系数， $S_{\hat{\beta}_i}$ 为样本方差。

3. F 检验

F 检验的目的是检验整个回归方程的回归系数是否合理。当参数检验都显著时，F 检验一定显著，但当 F 检验显著时，并不代表每个回归系数的 t 检验都显著。假设显著性水平为 α ，如果回归方程所求的 F 值满足 $F > F_\alpha$ ，则认为因变量和自变量有很显著的线性关系。

第3章 基于情感词典的网络舆情倾向性分类研究

针对微博文本倾向性的特点,本文借助情感词典对 SVM 算法中的特征选择算法进行改进,构建一个判别微博情感倾向的文本分类器。本章 3.1 节介绍基于 HowNet 的情感词典;3.2 节介绍基于情感词典框架下的网络舆情倾向性分类;3.3 节介绍实验测试及结果分析。

3.1 基于 HowNet 的情感词典构造

关于情感词典的构造,各个国家在自然语言处理方面都建立了相应的知识库,如普林斯顿大学的英文 WordNet^[45],微软的 Mindnet^[46],韩国的 Koreanwordnet 以及中国的 HowNet^[47]。英语词典相对中文词典有更深入的研究探索。Subasic 等^[48]人使用人工的方法建立了英语情感词典,该词典中的每一个词都标注了词的情感强度。较为著名的词典是 Wordnet,它是由普林斯顿大学的语言学家和计算机工程师共同开发设计的英语词典。而中文词典现在还在研究发展阶段,徐琳宏等^[49]人定义了中文词汇的倾向类别、强度的框架,并不断进行更新补充情感词库。其中,HowNet 是中文情感词典中最重要的词典之一。

3.1.1 HowNet 简介

HowNet (又名知网)是一个面向概念的知识库,并揭示概念之间和概念所具有属性之间的关系。它是一部十分详细的语义知识词典,它的特点是基于世界知识来构建语义网,采用网状结构来描述知识,而不是用树状结构来描述知识。HowNet 描述了以下几种关系:

1. 同义关系 (Synonyms):
2. 反义关系 (Antonyms):
3. 上下位关系 (Hypernym-Hyponym):
4. 对义关系 (Converse):
5. 部分-整体关系 (Part-whole):

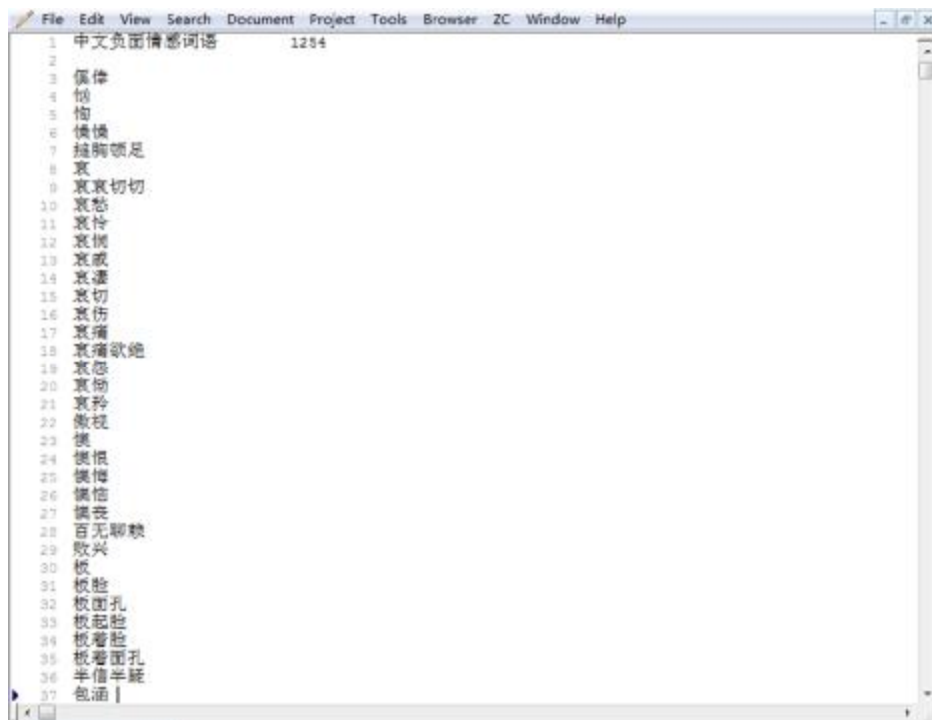


图 3.1 HowNet 部分中文负面情感词语

HowNet 包括两个重要的名词：概念和义原。概念可以定义为一种“知识表示语言”，这种“知识表示语言”是由“义原”构成的，而义原是知网中最小的、最基本的组成单位。也就是说，任何一个概念都可以由多个义原组合而成。在 HowNet2008 语料中，一共有 2000 多个义原，它们可以组成无限个概念集合。

3.1.2 基于 HowNet 的词语相似度计算

HowNet 中的任何概念都可以分解成若干个义原，刘群^[50]等学者提出了一种计算义原之间的相似度的方法，主要借助于义原树中的距离，最后两个词的相似度是以两个词之间的最大义原相似度为准， $\alpha (\alpha > 0)$ 为变化的参数，公式如下：

$$\text{similarity}(\text{word}_1, \text{word}_2) = \frac{a}{a + \text{distance}(\text{word}_1, \text{word}_2)} \dots\dots\dots (3.1)$$

但是这种方法有些时候并不是十分理想，比如“幸福”和“健康”的相似度是 0.76，而“幸福”和“悲剧”的相似度是 0.81。所以需要对上述方法进行改进，即在计算两个词相似度时，可根据该词和两组基准词（一组正向词，一组负向词）之间的相关程度，得到语义倾向性值。 $base_1$ 是正向的基准词， $base_2$ 是负向的基准词。则词 $word$ 的情感倾向为：

$$tendency(word) = \frac{1}{n} \sum_{i=1}^n similarity(word, base_1) - \frac{1}{m} \sum_{i=1}^m similarity(word, base_2) \quad \dots(3.2)$$

如果 $tendency(word)$ 为正, 说明 $word$ 为正向词, 如果 $tendency(word)$ 为负, 说明 $word$ 为负向词。而且, 朱妍岚等人实验验证了词 $word$ 的倾向性与选择的 $base_1$ 和 $base_2$ 有很大关系, 基准词越多, 其准确率就越高。

3.1.3 知网 HowNet 词典的扩展

由于本论文针对网络舆情的倾向性进行研究, 所以构建特定网络舆情主题的词典(也称领域词典)能够提高舆情分析的准确性。虽然极性词典很难达到通用的作用, 但是我们可以研究文本倾向性分类的基础上根据特定的主题不断填充和完善已有的通用情感极性词典。本论文所使用的构建方法中, 用二元组来定义一个词的所有属性, 其格式为: $WORD = (S, C)$, 其中 S 表示词语的情感极性, 该属性的取值为 1 或 -1 (1 代表正向的词语, -1 代表负向的词语), C 表示该词语的词性, 如名词 N 、形容词 A 、动词 V 等。而且本情感词典不包含任何没有情感倾向的词语。

1. HowNet 词典的优化精简

截止到 2007 年 10 月, 知网 HowNet 将词语划分为正向 $positive$ 和负向 $negative$ 两大类, 其中 $positive$ 词语 4566 个, $negative$ 词语 4370 个。在已经发布的词典中, 只是列出了词条和词典的总数, 词条其他信息(如一个词语是否为多义词等信息)则需要查询 HowNet 才能获得。

因为本论文所研究的对象是网络舆情, 所处理的文本都是网民常用的词语, 这些词语都是使用频率相对较高的, 而对于十分生僻的汉字很少使用甚至基本不用, 所以, 为了提高实验的高效率, 我们去掉了知网中一部分较为生僻的汉字和词语。最后整理好正向和负向的情感词。

2. 网络词汇的添加

HowNet 是一个通用的情感词典, 为了适合网络舆情的分析研究, 本文在(1)的基础上添加了网络中常用的情感词汇。共收录了 1583 个网络词汇, 具体格式为一个二元组: $NET = (S, W)$, 其中 S 表示词典中词条的情感极性 (1 或 -1), W 表示具体的词条。下表 3.1 是部分网络词汇。

表 3.1 部分网络词汇示例

情感极性	词条	词条解释
1	顶	“顶”表示赞同、认可的含义。如果某人在某论坛发表言论，其他人“顶”，就是将该言论置于最上方。
1	点赞	表示赞同或喜爱，体现一种心理认同。
1	给力	形容很给劲、很带劲的意思。
-1	拍砖	对某人所评论的观点不认同的评论或帖子。
-1	菜鸟	原指计算机水平较低的人，后引申为在某领域不拿手很差劲的人。
-1	抓狂	形容受不了某人所发帖子或评论而行为异常。

注：表中“词条解释”一列不存在词典中，只是在此为了说明解释。

网络词汇是不断更新的，每天都有可能产生新的网络词汇，因此网络词典也要不断的更新。而且一些网络词汇可能不规范，褒贬难辨，所以所有添加进网络词典的词条都采用小组投票的方式来决定其情感极性。

3. 网络表情符号的添加

在网络社交平台中，用户会经常使用表情符号表达自己的看法和态度，例如心情好比较高兴会使用“微笑”，失望糟糕或不同意某种观点会使用“伤心”等表情符号。下图 3.2 是部分网络表情符号。












表情图片	字符源代码	情感关键字
	😊	微笑
	😄	开心
	😊	美女
	😁	发呆
	😎	墨镜
	😓	哭
	😓	羞
	😓	睡
	😓	睡
	😓	哭
	😓	睡

图 3.2 网络表情符号示例

用户在网络媒介中,无论是聊天或是评论经常会使用表情符号表达自己的态度,陈述自己的观点。所以,表情符号在网络舆情分析中判断情感倾向是很重要的因素之一。虽然这些表情符号是一个个的动态图片,实际上在后台处理的过程中每一个符号都对应一组汉字,如图 3.2。这样,所有表情将转化为对文本的处理,大大地简化了问题的难度。表情符号的文本格式一般是用一个中括号来表示,比如:“😂”表示为“[哈哈]”。本论文使用微博表情统计的方法构建表情词典。

以微博为例,设微博的表情符号集合为 $S = \{s_1, s_2, s_3, \dots, s_n\}$, $s_i (i=1, 2, \dots, n)$ 为微博中的一个表情符号, 设微博的集合为 $WB = \{wb_1, wb_2, wb_3, \dots, wb_m\}$, 其中 wb_j 是一个二元组 $wb_j = \langle text, s \rangle$, wb_j 表示一条微博, $text$ 表示微博的文本, s 表示微博中的表情符号。对于每一个表情 s_i , 如果有 p 条微博包含 s_i , 则可以根据这 p 条表情的情感倾向来判断该表情的情感倾向, 进而构造表情符号情感词典。具体算法^[51]:

Step1: 提取微博发帖或评论中的表情符号。使用正则表达式或其他编程方法提取表情符号的文本形式, 构成集合 $HX = \{hx_1, hx_2, hx_3, \dots, hx_k\} (k < n)$;

Step2 : 计算一个表情符号 hx_i 的情感倾向。
 $hx_i \in WB$, $WB = \{wb_1, wb_2, wb_3, \dots, wb_p\} (p \leq m)$, 根据 $text$ 的文本倾向性判断 hx_i 的情感倾向 $E_{text_i} (=1|-1)$;

Step3: 计算 hx_i 的正向情感倾向值 $Q_{hx_i}^+$:

$$Q_{hx_i}^+ = \frac{\sum_{i=1}^p E_{text_i} = 1}{\sum_{i=1}^p |E_{text_i}|} \dots\dots\dots (3.3)$$

Step4: 计算 hx_i 的负向情感倾向值 $Q_{hx_i}^-$:

$$Q_{hx_i}^- = \frac{\sum_{i=1}^p E_{text_i} = -1}{\sum_{i=1}^p |E_{text_i}|} \dots\dots\dots (3.4)$$

Step5: 计算 hx_i 的综合情感倾向值 Q_{hx_i} :

$$Q_{(hx_i)} = Q_{(hx_i)}^+ + Q_{(hx_i)}^- \dots\dots\dots (3.5)$$

对于任何一个表情符号 hx_i , 如果 $Q_{hx_i} > 0$, 则表明 hx_i 的情感极性是正向的;

如果 $Q_{hx_i} < 0$ ，则表明 hx_i 的情感极性是负向的；如果 $Q_{hx_i} = 0$ ，则表明 hx_i 的情感极性是中性的。部分统计结果如下表 3.2:

表 3.2 网络表情符号及其情感极性

编号	表情符号	情感极性
1	 [嘻嘻]	正向
2	 [泪]	负向
3	 [怒]	负向
4	 [赞]	正向
⋮	⋮	⋮

4. 舆情词汇的添加

舆情词汇，顾名思义是与网络舆情相关的，能够表达一定的舆情情感倾向的词语。如：“看守所”、“监狱”、“检察院”等词语。我们采用小组投票的方式对舆情词汇进行情感极性的标注，分别为正向和负向。

3.2 基于情感词典框架下的网络舆情倾向性分类

网络舆情倾向性分类是对一段网络评论进行倾向性类别判断，判断该网络评论在某一观点上支持还是反对，属于二分类问题，属于传统的文本分类问题，所以传统的文本分类算法对于网络舆情倾向性分类是很重要的技术基础。同时，本文提出在情感词典框架下对倾向性分类算法中的特征选择方法进行改进，并对实验结果进行对比和分析。网络舆情的情感倾向性分类整体框架如下图 3.3:

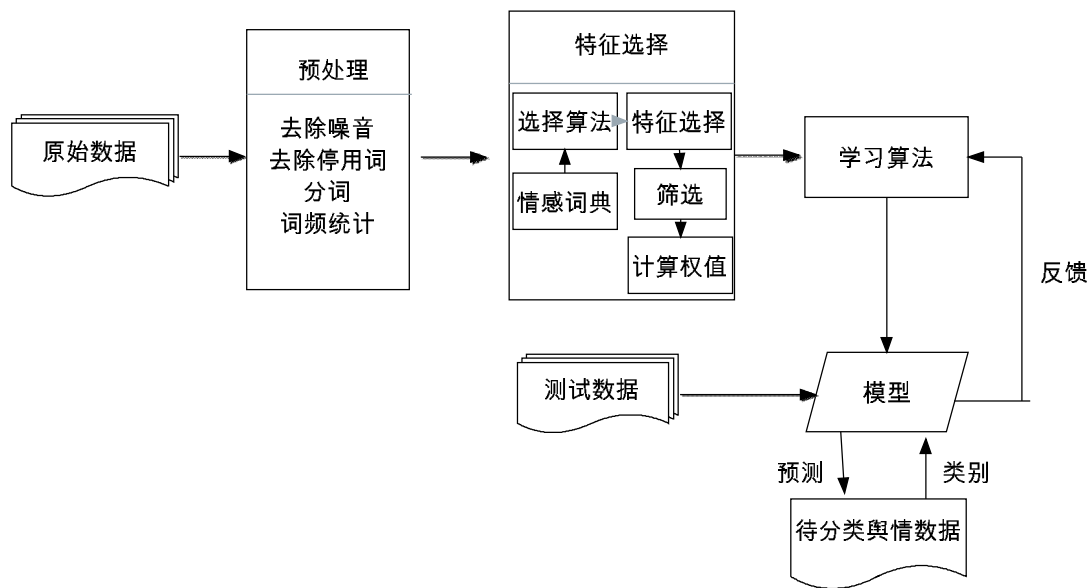


图 3.3 网络舆情倾向性分类整体框架图

3.2.1 数据预处理

本文的原始文本数据来自新浪微博,属于自然语言,是网民发表的个人观点,而计算机对于自然语言并不识别,所以需要将其转化为计算机能够识别的形式。在转化的过程中,要经历去噪声、分词、去除停用词等步骤。

去噪声: 去噪声是本文实验数据的初级过滤,由于本文使用的数据是新浪微博数据,微博数据一般是短文本,比如微博或 Twitter 其字数要求都在 140 个字以内,用户大多数必须使用精简的关键词语表达自己的观点。因此微博数据文本书写规范性没有那么严格,完整性较差。本文只对发帖的人的评论进行处理,不需要深入的挖掘深层信息,所以微博数据中会存在一定的噪声。比如:网址链接 (<http://t.cn/8FBnRRF>), 标点符号 (“【”、“】”、“#”、“@”、“:”等)。这些噪声对于实验没有任何价值,所以应当将其去除。

分词: 本文的实验数据是中文文本,文本的特点是由字构成词,由词和标点符号构成句,由句构成段、篇。需要将这些文本进行形式化处理,用一些特征来表示这些文本,再用这些特征在分类时区分文本。选择什么级别的粒度单位作为特征对于分类效果有很大的影响。选择细粒度的字作为特征,区分度不强,而且特征数量过多,影响分类效果;选择粗粒度的句或篇作为特征,特征过于复杂,同样达不到很好的效果。而词是介于二者之间的语言成分。所以,本文采用使用较为成熟的词语作为特征,避免了特征粒度过大或过小带来的弊端。但是,中文

文本相对于英文文本来说有一个困难,英文的词是通过空格来区分的,很好处理,而中文词与词之间没有特殊的标识,它们都是连在一起的,计算机无法识别词语,这就需要使用分词工具解决这个问题。下面是一些常用的分词方法:

1. IKAnalyzer。IKAnalyzer 是一个基于 java 语言的开源的中文分词工具包。从第一个版本发布后,至今共发布 3 个版本。IKAnalyzer 在初期是以开源项目 Luence^[52]为主体的,主要使用词典分词和文法分析算法的中文分词组件。后来随着 IKAnalyzer 的不断发展,尤其是 IKAnalyzer3.0 版本采用“正向迭代最细粒度切分算法”,每秒能处理六十万字,该版本慢慢脱离 Lucene 项目,主要面向 java 的公共分词组件。同时,该分词工具采用了子处理器分析模式,能够对英文字母、中文词汇(姓名、地名)、数字等进行处理。而且它优化了词典存储占用很小内存,同时还支持用户词典扩展,用户可以根据特殊需求扩展词典,方便用户灵活使用。

2. ICTCLAS。ICTCLAS 是中国科学院计算技术研究所开发出的汉语语法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System),它全部使用 C/C++语言编写,同时支持 Linux、Windows 和 FreeBSD 等操作系统。ICTCLAS 主要包括中文分词、命名实体识别、词性标注、新词识别和用户词典等功能。目前最新版本 ICTCLAS3.0 分词精度达到 98.45%,是比较优秀的一个中文分类器。

3. LibMMSEg。LibMMSEg 是 Coreseek.com 设计的中文分词软件包,其主要使用的算法是 Chih-Hao Tsai 的 MMSEG,其分词过程是在 GPL 协议下进行的。LibMMSEg 的编程语言是 C++,同时支持 Linux 和 Windows 操作系统,切分速度大约每秒钟 30 万,在切分速度方面仍有较大提升空间。

去除停用词:停用词(stop words),词典解释为“电脑检索中的虚字或者非检索字”,最初是在搜索引擎中使用,当索引或处理搜索词时自动忽略的那些词被称为停用词。一般来说,停用词的特点是在文本中经常出现,但是意义不大。包括副词、介词、连词等,如常见的“了”、“的”、“使”、“其中”等词语,都属于停用词。如果文本中存在大量的停用词,会对最终处理结果造成一定的干扰,所以在文本处理时需去除停用词。IKAnalyzer 分词工具本身具有去除停用词的作用,在原来停用词表基础上,本文进行扩充,使用比较完整的停用词表,共有

1962 个停用词，包括中文、数字、标点、特殊符号等。

本文使用 IKAnalyzer 分词工具，举例：“为马航每一条生命祈祷，愿你们平安”，分词的结果是“马航 一条 生命 祈祷 愿 你们 平安”。

3.2.2 文本特征提取

经过上述的预处理过程，得到的词都是具有一定重要程度的词语。在文本数据的形式化表示中，所有特征构成特征空间，每个文档可以用多个特征来表示，每个特征表示一个属性或维度，如果有上万条的微博数据，其特征空间的维度将十分巨大，导致维度灾难，影响文本倾向性分类效果。

特征选择 (Feature Selection) 也叫作特征子集选择，是从上述中庞大的特征中选择一个特征子集，来构造实验模型。特征选择能够去除不相关的特征，冗余的特征，从而达到降维的效果，减少运行时间的目的。特征选择的任务就是对特征进行量化，突出表示不同特征的重要性。下面简单介绍三种较为成熟的特征选择算法：

1. 文档频率方法 (Document Frequency)

文档频数是指在所有数据集中包含该词的文档数，是一种很简单的特征选择算法。在训练数据时，对于每一个特征都计算它的文档频数，如果这个频数大于或小于预先设定的阈值，则将该特征删除。因为如果一个特征的频数过小会导致该特征“没有代表性”，如果该特征的频数过大则会导致该特征“没有区分度”，所以合理选择阈值十分重要。该方法的优势是时间开销小，速度快，计算每个特征频数的时间复杂度为线性的，比较适合大规模数据的特征选取。所以文档频率 DF 是一种比较简单可行的特征提取方法。

但是在实际应用中，并不会直接使用 DF，而是把它作为评判其他方法的一个标准。下面给出一个特征 f 的文档频率 $DF(f)$ 计算公式：

$$DF_f = N_f / Sum \dots\dots\dots(3.6)$$

其中，Sum 表示数据集中文档的总数， N_f 表示 Sum 中包含特征 f 的文档数量。

2. 卡方检验

卡方检验主要是用于衡量统计观察实际值和理论推断值之间的偏离程度，卡

方值的大小反应了实际观测值和理论值的偏离大小。卡方值越大,结果越不符合条件;卡方值越小,结果越符合条件。如果卡方值足够小,则可以认为是误差造成的,接受原假设;否则接受备择假设。

卡方检验的主要目的是衡量词条 t 和文档类别 c 之间的相关程度,如果二者服从自由度为一阶的卡方 (CHI) 分布,那么 t 对于 c 的卡方值的计算公式:

$$c^2(t,c) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \dots\dots\dots(3.7)$$

其中, N 为数据集的文档总数, A 为包含 t 而且属于 c 类的文档总数, B 为包含 t 但不属于 c 类的文档总数, C 为不包含 t 但属于 c 类的文档总数, D 为不包含 t 而且不属于 c 类的文档总数。文献^[53]在此基础上又做了改进,提出基于方差的 CHI 特性选择方法:

$$c^2(t,c) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \times p \times q \times r \dots\dots\dots(3.8)$$

其中, p 主要解决 CHI 统计方法对文档低频词缺陷的问题, q 用来衡量一个类别中含有特征词的文档频偏离全部类别文档频综合值的程度, r 用来衡量一个文档的词频偏离全部文档词频综合值的程度。

(3) 信息增益^[9] (IG, Information Gain)

一般认为,如果一个特征包含的信息量越多,那么该特征也就越重要,该特征就应该赋予更高的权值,通常用“熵”来表示特征包含信息量的多少。IG 就是一种基于熵的评估方法,定义某个特征所含信息量,它通过训练数据计算出每个特征所包含的信息增益,然后按照增益值从大到小排序,选择出排名靠前特征。信息增益选择的特征属性只能反应一个特征对整个系统的贡献,却不能具体到某个类别上,所以信息增益只能做全局性的特征选择,即所有的类使用相同的特征集合。

3.2.3 线性组合特征选取算法

预处理后的数据仍是文本形式的,计算机并不能识别这种格式的数据,需要将文本形式的数据转化为向量形式的格式。学术界普遍使用的是向量空间模型 (VSM, Vector Space Model), 主要将文本内容转化为向量空间中的向量运算,通过计算向量的相似度表达反映语义的相似度。在文本处理中, VSM 将每个文

档表示成多维 (n 个) 特征的一个向量, 如果数据集中含有 m 个文档, 那么整个数据集可以表示成一个 $m \times n$ 维的向量空间, 完成了从文本到向量的抽象化过程^[54]。下表 3.3 形象的表达了文本转化为向量的形式, 其中 $w_{i,j}$ 表示权值。

表 3.3 使用向量表示文档

	feature₁	feature₂	...	feature_n
doc₁	w_{11}	w_{12}	...	w_{1n}
doc₂	w_{21}	w_{22}	...	w_{2n}
...	w_{ij}	...
doc_m	w_{m1}	w_{m2}	...	w_{mn}

但是, 文本数据与其他图像等数据不同, 即使经过预处理剩下的词条也有几万、几十万条。如果把这样的文本转化为向量空间模型, 其特征空间也是十分巨大的, 尤其是微博数据, 每条数据字数有限, 数据条数又多, 必然导致数据稀疏的问题。为了获得理想的实验效果, 减少高维数据带来的麻烦, 需要选取合理的特征, 降低特征的维度。文本针对特征选择, 提出了线性组合的特征选择算法 (LC-FS, Linear Combination – Feature Selection), 主要是改进的两种特征选择方法的结合: 词频方法和互信息方法。

下面是具体的特征选择过程:

1. 词频方法

词频也是衡量特征重要性的一种方法, 根据对预处理分词后的数据结果, 统计所有词的频率 $F(t)$, 将频率为 1 的词条 (特征) 删除, 剩下的词条按照词频从大到小排序。词频越大, 表明该词越重要。

2. 互信息方法

互信息是信息论中的概念, 也是特征选择中一个重要的算法, 描述了两件事情发生时相关联相互影响的信息量。例如, 在处理本文微博文本分类的特征提取时, 可以用互信息来衡量一个特征和特定类别 (正向或负向) 的相关性。如果该特征和类别的互信息越大, 则二者相关联的信息量就越大, 该特征和这个类别的相关性就越大。

由于本文处理的对象的是网络舆情文本信息, 所以还要根据网络舆情的具体特点进行研究, 下面将详细来叙述:

由本章 3.1 可以得到针对网络舆情的感情极性词典, 该词典包括 HowNet 基础词典、网络新词词典、表情符号词典、舆情词汇等。因为本文研究的是网络舆情分析, 网络舆情有正面的、负面的消息, 而网络舆情往往更加关注事件的负面消息, 是否会对社会和人民带来隐患, 所以为了更好的解决这个问题, 构建的感情极性词典也应该适合网络舆情分析。

假设感情极性词典集合为 $S = PD \cup ND$ (PD 为正向的词语集合, ND 为负向的词语集合), 数据集词语频率大于 1 的集合为 D , 则最后的候选词集合为 $H = S \cap D$, 即感情极性词典和数据集共同存在的词, 其个数为 n 。接下来, 使用互信息算法进行特征选择, 计算集合 H 中的词条 t 与类别 c 的互信息量, 将计算结果按照互信息量从大到小排序, 词条 t 与类别 c 的互信息计算公式为:

$$I(t, c) = \log_2 P(t/c) / P(t) \dots\dots\dots (3.9)$$

其中, $P(t|c)$ 为在属于类别 c 的数据中存在词条 t 的概率, $P(t)$ 为词条 t 所在的文档数与总文档数之比。为了计算词条 t 在全局词条中的作用, 计算词条关于特征的平均值 $I(t)$:

$$I(t) = \sum_{i=1}^k P(c_i) I(t, c_i) \dots\dots\dots (3.10)$$

其中, $P(c_i)$ 为类别 c_i 的概率, k 为类别的数量。

3. 线性组合特征选择

在做特征选择的过程中, 互信息的度量标准是以特征项在不同类别中出现的次数在整个数据集中的比值来决定特征属于哪一类别。如果在类别 c_i 中出现的频率高, 而且在训练集中出现的频率低的特征 t_k 对 c_i 类别的引导能力就越强。所以, 互信息的优点是考虑了低频词所携带的信息量。互信息方法正是这样一个特点, 也带来了一定的不足: 首先, 它放大了低频词的作用, 对低频词过于敏感; 其次, 它没有考虑分别不均匀的训练样本类别对特征选择结果的影响。正是存在这样的不足, 本文提出将互信息和词频相结合, 弥补互信息存在的缺陷。下面给出线性组合特征选择的公式定义:

定义: 根据线性相关的定义, 本文将互信息和词频相结合进行特征选择, 定义如下:

$$Score_{LC-FC} = I Freq(t_i) + (1 - I) I(t_i) \dots\dots\dots (3.11)$$

其中, $Score_{LC-FC}$ 为词条 t_i 的得分, $Freq(t_i)$ 表示词条 t_i 的词频, $I(t_i)$ 表示词条 t_i 的互信息值, $\lambda (0 \leq \lambda \leq 1)$ 是调节因子。由于词频 $Freq(t_i)$ 和互信息值 $I(t_i)$ 的度量单位不同, 需要将其进行特殊处理。这里, $Freq(t_i)$ 和 $I(t_i)$ 都经过归一化处理, 即:

$$Freq(t_i) = \frac{F(t_i)}{\max\{F(t_i)\}} \dots\dots\dots (3.12)$$

$$I(t_i) = \frac{IN(t_i)}{\max\{IN(t_i)\}} \dots\dots\dots (3.13)$$

其中, $F(t_i)$ 是词条 t_i 初始的词频值, $IN(t_i)$ 是词条 t_i 初始的互信息值, 总的词条数为 n 。

对于公式(3.11), 当 $\lambda=1$ 时, $Score_{LC-FC}$ 值回归为简单的词频特征选择方法; 当 $\lambda=0$ 时, $Score_{LC-FC}$ 值回归为互信息方法; 当 $0 < \lambda < 1$ 时, $Score_{LC-FC}$ 优化两种方法, 构造线性组合的形式, 避免二者的不足之处。根据公式 3.11 计算出所有词条 t 的 $Score_{LC-FC}$ 值, 然后按照从大到小的顺序排序, 根据经验事先设定一个特征数目阈值 θ , 将排在阈值 θ 前面的那些特征抽取出来作为最后的特征。下面给出线性组合特征选择算法 LC-FS 的具体过程:

算法: LC-FS

输入: 分词预处理之后的数据集合 SE , 情感词典 $S = PD \cup ND$, 特征个数 θ

输出: 特征集 F

1. 统计集合 SE 中每个词的词频 $F(t_i)$
2. **For** 每个词 t_i
3. **If** $F(t_i) > 1$
4. 加入到集合 D 中
5. **End For**
6. 根据 $F(t_i)$ 从大到小排序
7. **For** 每个词 $word_i \in D$
8. **If** $word_i \in S$
9. 加入到集合 H 中

-
10. **End For**
 11. **For** 每个 H 中的词条 t_i
 12. 计算词条 t_i 与类别 c 的互信息
 13. 计算词条 t_i 的平均互信息值 $I(t_i)$
 14. **End For**
 15. **For** $\lambda \leftarrow 0$ to 1 // λ 取值间隔为 0.1, 共 11 个数
 16. **For** 每个 H 中的特征
 17. 计算每个词条的 $Score_{LC-FC}$ 值
 18. **End for**
 19. 计算每个 λ 对应的 $Score_{avg_\lambda}$ 值
 20. **End For**
 21. 根据 $Score_{avg_\lambda}$ 值, 取最大的 $Score_{avg_\lambda}$ 值对应的 λ 值, 并将该 λ 作为后续实验的参数
 22. 依据 $Score_{LC-FC}$ 值从大到小排序, 选择前 θ 个特征加入特征集合 F
 23. **Return** F
-

3.2.4 特征权重计算

特征权值是通过计算一个或多个评价指标, 实现对特征空间中的特征进行量化。在向量空间模型下, $TF-IDF^{[41][55]}$ 是比较著名且常见的一种文本特征赋值方法。 $TF-IDF$ 主要考虑了两个重要因素: 一个是词频 TF (Term Frequency), 另一个是逆向文档频率 (Inverse Document Frequency)。按照 $TF-IDF$ 的理论, 如果一个词在一个文档中出现的次数很多, 同时在其他文档中出现的次数很少, 那么这个词具有一定的文档区分能力。

词频 TF : 即词语在文档中出现的次数。如果一个词在一篇文档中重复出现, 那么它就具有了表征该文档的能力。一般认为, 一个词语在文档中的重要程度和词语的出现频率是正相关的。这里的 TF 是对词频数的归一化结果, 为了避免它偏向篇幅长的文档。则文档 d_j 中的词 t_i 的词频 tf_{ij} 为

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \dots\dots\dots (3.14)$$

其中, n_{ij} 表示词 t_i 在文档 d_j 中出现的次数, $\sum_k n_{kj}$ 表示文档 d_j 中其它所有词出现的次数之和。

逆向文档频率 **IDF**: **IDF** 是一个全局的度量标准, 考察特征之间的重要程度。如果一个词语在很多文档中都出现, 那么该单词就可能在多个类别的文档中出现, 其结果就是它不具有区分一种类别的能力, 即便它在一个文档中出现的频率再高也不能很好的区分能力。一般认为, 一个词是否具有类别区分能力与逆向文档频率也是正相关的。下面列出词 t_i 的逆向文档频率 idf_i :

$$idf_i = \log \frac{|DOC|}{|\{j: t_i \in d_j\}|} \dots\dots\dots (3.15)$$

其中, DOC 为数据集中的文档总数, $|\{j: t_i \in d_j\}|$ 表示所有文档中包含词 t_i 的文档数目。

基于上述描述, 特征的重要程度与词频 **TF** 和逆向文档频率 **IDF** 都成正比。所以, **TF-IDF** 特征权值 w_{ij} 的计算公式表示为:

$$w_{ij} = tf_{ij} * idf_i \dots\dots\dots (3.16)$$

但是, 在实际使用过程中, 上述公式(3.15)(3.16)存在一定的缺陷, 需要对其加以改进: idf_i 中的分母可能为 0, w 值的归一化处理。改进后的计算公式为:

$$w_{ij} = \frac{tf_{ij} * \log(\frac{|DOC|}{|\{k: t_i \in d_k\}|} + a)}{\sqrt{\sum_{i=1}^n \left[tf_{ij} * \log(\frac{|DOC|}{|\{k: t_i \in d_k\}|} + a) \right]^2}} \dots\dots\dots (3.17)$$

其中, w_{ij} 为第 j 个文档的第 i 个特征的权值, n 为特征的总数, $k \in (1, |DOC|)$, a 为一个很小的数但不等于 0。

经过以上处理, 完成了将非结构化的自然语言到计算机可以识别的结构化的向量形式的转变, 下面将介绍模型的训练实验过程。

3.3 实验测试及结果分析

3.3.1 实验环境及参数选取

本文的实验环境：CPU 为 E5700 3.00GHz，内存为 2GB，OS 为 Win7，以及基于 java 编程语言在软件平台 Eclipse 的环境下进行实验。

本文的数据来源于新浪微博马航飞机失事事件，共包括 24734 条微博用户评论信息，时间跨度为 2014 年 3 月 1 日到 2014 年 4 月 10 日，其数据的属性列有用户 ID、发帖时间、发帖内容、相关网址等信息。而且，该数据的文本倾向性都是通过人工标注的，类别包括正向（标记为“1”）和负向（标记为“-1”），即解决二分类问题。同时，随机选择 14734 条微博数据作为训练集，其余 10000 条微博数据作为测试集。随机选择是为了保证无偏见性以及后续实验的交叉验证。下表 3.4 是部分数据的格式：

表 3.4 原始数据的格式

编号	微博内容	时间	极性
1	马航出内鬼，可悲……我分享了 http://t.cn/8sVJ3xu	2014-04-04	-1
2	马航飞机背后也许有着更大的阴谋， 但乘客是最大的受害着！	2014-03-21	-1
3	#马航飞机失联# 是个阴谋，以阴谋之心 度阴谋之举	2014-03-17	-1
4	#祈福马航# 希望所有人平安，为你 们祈福 [蜡烛][蜡烛][蜡烛][蜡烛]	2014-03-08	1
...

3.3.2 分类器学习方法的选择

支持向量机（SVM）是监督式学习方法中比较重要的一种学习方法，适合大样本集的分类，尤其是文本分类。由于 SVM 具有以下特点：（1）是一种由坚实基础理论的小样本学习方法，它不涉及概率测度和大数定律，不需要太多的训练数据；（2）它的决策函数由支持向量的个数决定，且计算复杂度与支持向量的

个数有关,避免了维度灾难;(3)支持向量样本集有较好的鲁棒性;(4)它对核函数的选取不敏感。所以,本文选取 SVM 方法作为分类器的学习方法,实验使用台湾大学林智仁教授开发的 Libsvm 来实现 SVM 算法。在文本转化成向量的基础上,还需要进一步转化为 Libsvm 的数据输入格式:

$$\begin{bmatrix} \text{label} & \text{index}_1:\text{weight}_1 & \text{index}_2:\text{weight}_2 & \cdots \\ \text{label} & \text{index}_1:\text{weight}_1 & \text{index}_2:\text{weight}_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

其中,该数据格式是一个矩阵的形式,一行代表一个文本,列代表文本的数量,第一列 *label* 表示文本的类别,本文 $\text{label} \in \{1, -1\}$, 其余各列表示每个文本的特征, *index* 表示特征的标号, *weight* 表示对应特征 *index* 的权重。

SVM 核函数使用 RBF, 在参数 *c* 和 *g* 的选取方面, 首先使用粗略选择, *c* 和 *g* 的变化范围是 $(2^{-10}, 2^{-9}, 2^{-8}, \dots, 2^{10})$; 然后使用精细选择, *c* 和 *g* 的变化范围是 $(2^{-4}, 2^{-3.5}, 2^{-3}, \dots, 2^4)$, 其他所有参数使用默认值。

3.3.3 评估标准

查准率 (Precision) 和查全率 (Recall) 是文本分类中常用的评估标准, 能够衡量文本分类器的性能。为了清晰的表述二者的联系与区别, 下面使用二者的混合矩阵 (Confusion Matrix) 来表示:

表 3.5 混合矩阵

	被分为类别相关	被分为类别无关
实际与类别相关	TP	FN
实际与类别无关	FP	TN

$$\text{Precision} = \frac{TP}{TP+FP} \dots\dots\dots(3.18)$$

$$\text{Recall} = \frac{TP}{TP+FN} \dots\dots\dots(3.19)$$

Precision 表示被正确分类为与类别相关的文档数量除以被分类为与类别相关的文档数量, Recall 表示被正确分类为与类别相关的文档数量除以测试集中实际与类别相关的文档数量。对于一个分类器来讲, 如果单纯的使用这两个标准进行衡量, 其准确率可能很低, 但查全率可能很高。所以, 在实际应用过程中, 我们引入另一个评估函数 F-score, 它是由查准率和查全率决定的但又不完全由其

中一个来决定, F-score 是查全率和查准率的一个调和平均值, 综合考虑两者的因素, 只有**Precision**和**Recall**的值都高, 其评估标准 F-score 值才能高。

$$F - score = \frac{2 * Precision * Recall}{Precision + Recall} \dots\dots\dots(3.20)$$

3.3.4 算法测试及对比分析

下图 3.4 是微博网民的参与情况, 舆情刚刚发生时人们的参与程度相对密集, 随着时间的发展, 舆情参与程度逐渐变淡趋于稳定。

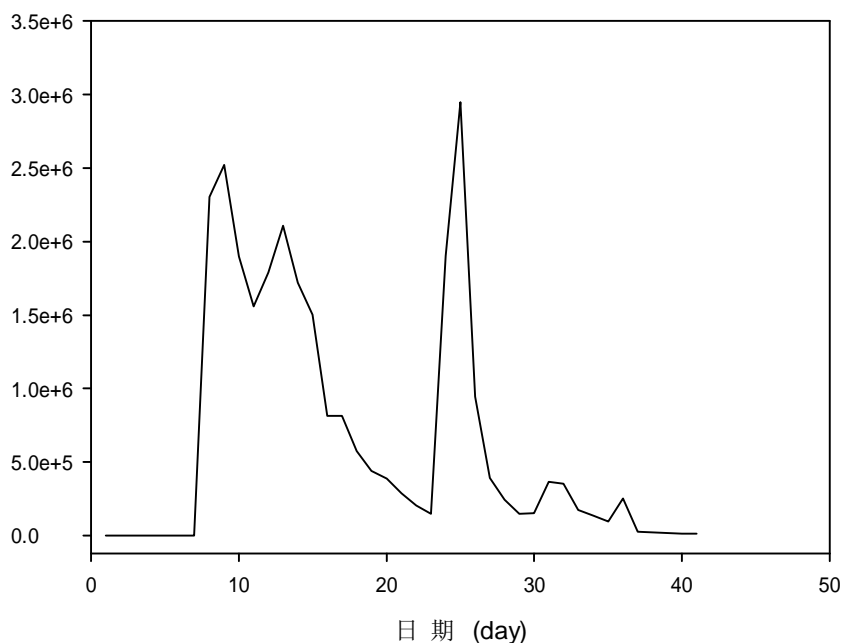


图 3.4 微博网民的参与程度

本章基于上述的数据进行实验, 一共分四组实验, 具体实验方案如下表 3.6, 四组实验的数据完全相同, 最后讨论分析各个方法对实验结果的影响。

表 3.6 实验方案

实验编号	不同实验方法
1	词频
2	互信息
3	情感词典+互信息
4	线性组合特征选择 (LC-FS)

实验1使用预处理后的数据,统计每个词的词频 $F(t)$,根据词频从大到小排序,按照事先设定的阈值选择词频靠前的词语作为特征,构成特征集合。此时,文档数据可以表示成 (d_1, d_2, \dots, d_n) ,其中 $d_i (i = 1, 2, \dots, n)$ 是特征词,然后使用权值计算方法 **TFIDF** 对每个文档计算相应词语的权值,将文档表示为向量空间的形式。

实验2在特征选择方面,通过计算每个词的互信息值 $IN(t)$,根据互信息值从大到小排序,同样根据事先设定的阈值,将满足条件(大于阈值)的词选择出来构成特征集合,后续的方法与实验1相同。

实验3引入情感词典的思想,同时结合互信息方法,将其应用在特征提取上。首先,求情感词典和分词后的词的交集;然后,对交集集中的每一个词计算与类别 c 的互信息值,按照设定的阈值选择满足条件的词作为特征,后面的权值计算等过程与实验1相同。

实验4在实验3的基础上,弥补实验3的不足(即互信息突出低频词的作用),将实验1的词频特征选择方法和基于词典的互信息计算方法相结合,引入因子 λ 综合两种方法的利弊,计算每个词的得分 $score$,再根据 $score$ 从大到小排序,选择满足条件数量的特征,后续方法与实验1相同。

下图3.5是线性组合特征选择实验方法所对应的带权值的文档向量表示。

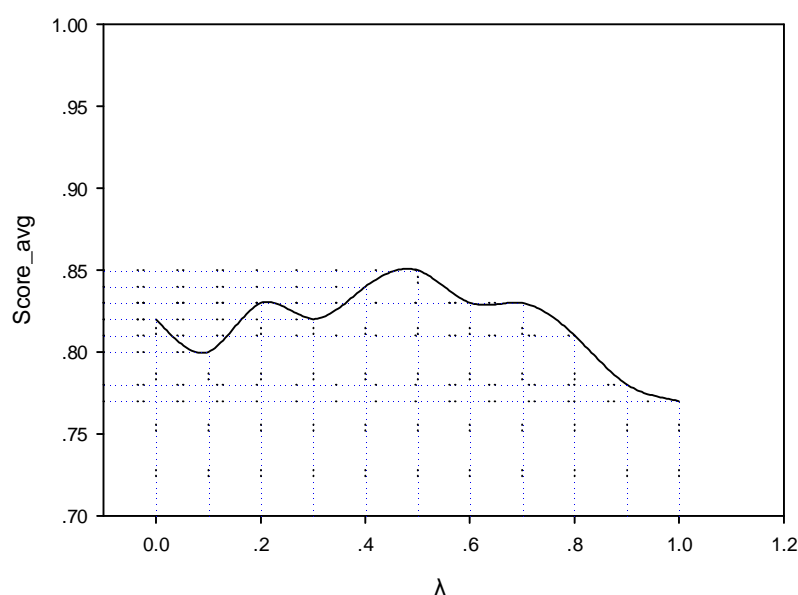
```
1 24:0.052357685 84:0.1388397 86:0.069461614 93:0.06988312 184:0.07837905
289:0.17448941 440:0.09556159 537:0.099905156 611:0.10329513 712:0.10722769
818:0.110542335 956:0.11543272 1204:0.12189095 1269:0.12296917 1379:0.25137088
1836:0.40339553 2064:0.13758524 3120:0.15152735 3456:0.15532185 4036:0.32301617
```

图3.5 LC-FS 特征选择方法的向量表示

在实验4中,由于 $Score_{LC-FS}$ 方法需要调整 λ 的值,选择效果最优的 λ 值,本实验 λ 值选择 11 个数值 (0, 0.1, 0.2, ..., 0.9, 1), 每个 λ 的平均值 $Score_{avg\lambda}$ 为

$$Score_{avg\lambda} = \frac{1}{n} \sum_{i=1}^n Score_{LC-FS_i}, n \text{ 为词条数} \dots\dots\dots (3.21)$$

由公式 (3.21) 可以计算得到每个 λ 的平均值,下图表述了 λ 的平均值的大小关系,从图 3.6 可以看出,当 $\lambda=0.5$ 时, $Score_{avg\lambda}$ 最高,即实验效果最好,所以选择 $\lambda=0.5$ 进行下面的实验。

图 3.6 λ 的不同取值

向量化的每个文档都可以用特征进行表示,而不同的特征个数对于文档的表征能力是不同的。为了找出最佳的特征个数,我们设定了 4 组特征个数的实验: 4500、5000、5500、6000。同时,这些实验都是基于互信息特征选择方法进行的,以下不同特征个数对实验的影响情况。

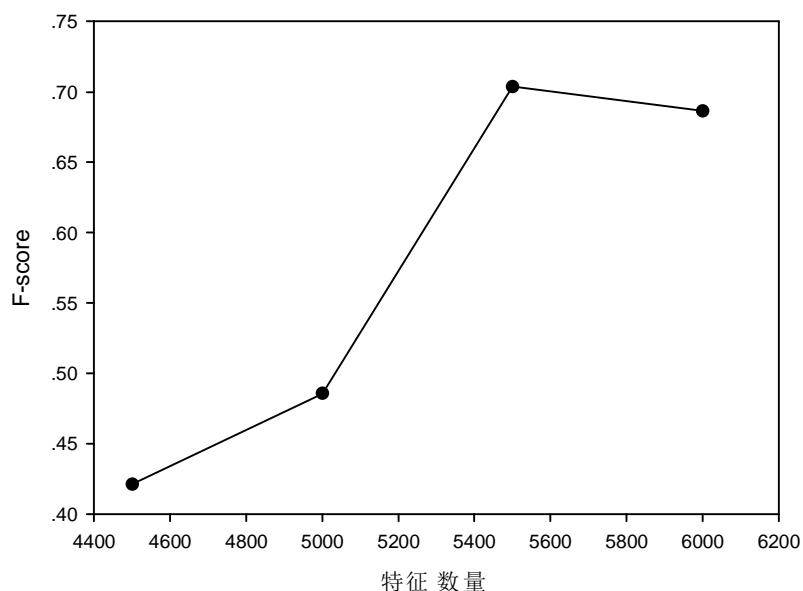


图 3.7 不同特征数量的分类效果比较

从图 3.7 中可以看出,当特征数量为 5500 时,实验的分类效果最佳。当特征过少或特征数量增多时,都不会得到较好的结果。所以,本文接下来的实验都

选取特征数量为 5500 为基础进行。

为了便于实验对比, 本文从正向舆情和负向舆情两个角度出发, 分别求出词频、互信息、情感词典+互信息、线性组合特征选择 (LC-FS) 四种方法对应的 Precision 值、Recall 值和 F-score 值。表 3.7 是本文的实验结果。

表 3.7 不同特征选择方法的实验结果

实验方法	正向舆情			负向舆情		
	Precision	Recall	F-score	Precision	Recall	F-score
词频	60.12%	52.47%	56.03%	55.44%	48.25%	51.60%
互信息	70.73%	66.54%	68.57%	61.34%	60.66%	74.42%
情感词典 +互信息	77.56%	69.35%	73.23%	72.89%	68.87%	70.82%
线性组合特征 选择 (LC-FS)	81.43%	75.09%	78.13%	79.37%	72.55%	75.80%

从上面表 3.7 的实验结果可以总结以下几点:

- 1、本文提出的线性组合特征选择 (LC-FS) 算法, 综合实验效果较好, 无论是正向舆情文本还是负向舆情文本, 其 F-score 值较其他方法都高。使用词频的特征选择方法效果最差。
- 2、由实验 2 和实验 3 可知, 本文构建的情感词典, 与互信息相结合, 在特征选择方面有显著的提高。说明情感词典在网络舆情情感倾向性分类中起到很大的作用。
- 3、由实验 3 和实验 4 可知, 线性组合特征选择 (LC-FS) 算法衡量了词频和互信息的优缺点, 选择质量更高的特征进行模型训练, 而且使用适合网络舆情的领域情感词典, 所以在四种方法中表现最佳。但整体 F-score 值不是很高, 可能是因为微博数据的短文本特点以及微博语言的不规范性造成的。
- 4、Precision 和 Recall 相互制约, 在保证 Precision 高的情况下, Recall 会相应的下降, 反之亦然。

第4章 时序信息的网络舆情演化规律模型

4.1 网络舆情演化

根据中国互联网网络信息中心发布的数据,到2015年6月,我国微博用户规模达到2.04亿,可见人们在网络平台进行交流意见、发表评论的人数逐渐增多。微博舆情已经成为社会舆情的新的领域延伸,随着微博用户的数量不断递增,微博对社会的影响力也不断增强。这样众多的网民在网络上每天发表不计其数的评论,所以掌握网络舆情的发展状态,对于政府和国家具有重要的意义。

网络舆情的演化对于研究网络舆情分析具有重要的意义,也是网络舆情领域的重要组成部分,可以从不同的维度对网络舆情的演化进行研究。为了方便研究网络舆情的演化过程,本文将网络舆情的演化过程分为三个部分:初始萌芽阶段、迅速扩散阶段和消退稳定阶段。同时,对各个阶段进行分析。下面详细介绍网络舆情演化的三个过程:

(1) 初始萌芽阶段

在初始阶段,只有少数的网民接触到舆情信息,获取的渠道可能包括新闻、电视、微博或是人们的口口相传,而且此时网民对舆情的认识程度还不够,对舆情的整个事件还不是特别了解,所以这个阶段网络的舆情热度还不是很高涨。但是这个阶段已经有人在网络平台(如微博)发表自己的观点,还没有形成认可度较高的群体观点。

(2) 迅速扩散阶段

网络舆情的信息越来越丰富,信息报道方式更加多样,网民对舆情已经有了基本充分的了解,网民积极发表个人的见解,包括支持的、反对的、中立的态度,而且每种态度的倾向性更加明显,网民通过各种信息获取方式对舆情的认识更加深刻,评论见解更加有针对性,群体逐渐形成强有力的、认可度较高的观点。网民之间慢慢形成群体极化现象。

(3) 消退稳定阶段

待到网络舆情发展到一定程度或其主体得到相应的处理之后,舆情本身的发展态势基本定型,网民对问题的辩论就会渐淡下来,网民的关注度逐步下降,舆

情接近尾声，直至趋于稳定，甚至舆情消失。

4.2 网络舆情演化的驱动因素

4.2.1 网络舆情驱动因素介绍

网络舆情之所以能够在上述三个阶段不断的传播，是受多种因素的驱动，网络信息的传播是一个复杂多变的过程。如果一个舆情事件发展的态势过于高涨，甚至逐渐具有威胁社会安全的趋势，那么相关部门应及时作出相应的举措进行解决。为了能够及时了解网络舆情的发展趋势，本文针对该问题综合考虑多个舆情驱动因素进行分析，并且重点关注负面舆情造成的影响。

针对网络舆情的具体事件，主要考虑以下几个方面，分别为：迁移指数、发帖热度、环境热度、搜索热度。这几个方面在网络舆情中从不同的维度影响着舆情的发展。

（1）迁移指数

迁移指数描述网络中某一主题在不同时间段之间的活跃程度。在时间段 $T = \{t_1, t_2, t_i, \dots, t_{n-1}, t_n\}$ ， t_i 代表不同的时间点，假定在 t_i 时间点的参与度为 p_i ， t_{i+1} 时间点的参与度为 p_{i+1} 。如果 p_i 与 p_{i+1} 相差很大，说明在 t_{i+1} 有新的主题出现或突然讨论减弱；如果 p_i 与 p_{i+1} 比较接近，说明讨论较平静或维持在激烈讨论之中。

（2）发帖热度

发帖热度主要指网民在单位时间内（一小时、一天、一周等）的发帖数量或评论回复数量，反应了网民对网络舆情的参与程度。以微博为例，微博发帖热度表示微博用户在一天内的参与某一主题的程度，包括发帖、转发、点赞、回复等一系列操作。

（3）环境热度

主要指舆情发生后的周围网络环境和社会环境的参与度，其中网络环境包括一些门户新闻网站等媒体，社会环境包括各大新闻报纸的报道，以及政府对舆情信息的相应态度等。可以使用这些线上线下的媒体对网络舆情的报道新闻的数量来描述环境热度。

（4）搜索热度

当网络中出现舆情信息时,网民通常会使用搜索引擎获取网络舆情,了解最新动态,把网民搜索的参与度进行量化称做搜索热度。搜索热度能够反映某个关键词在搜索引擎的搜索规模有多大,以及舆情的发展态势和相关新闻舆论变化。而且,通过搜索热度也能知道关注这些舆情的都是什么样的人,搜索的时间段,这对于研究网络舆情的演化规律具有重要的指导意义。

4.2.2 网络舆情驱动因素分析

(1) 网络舆情热度与驱动因素的关系

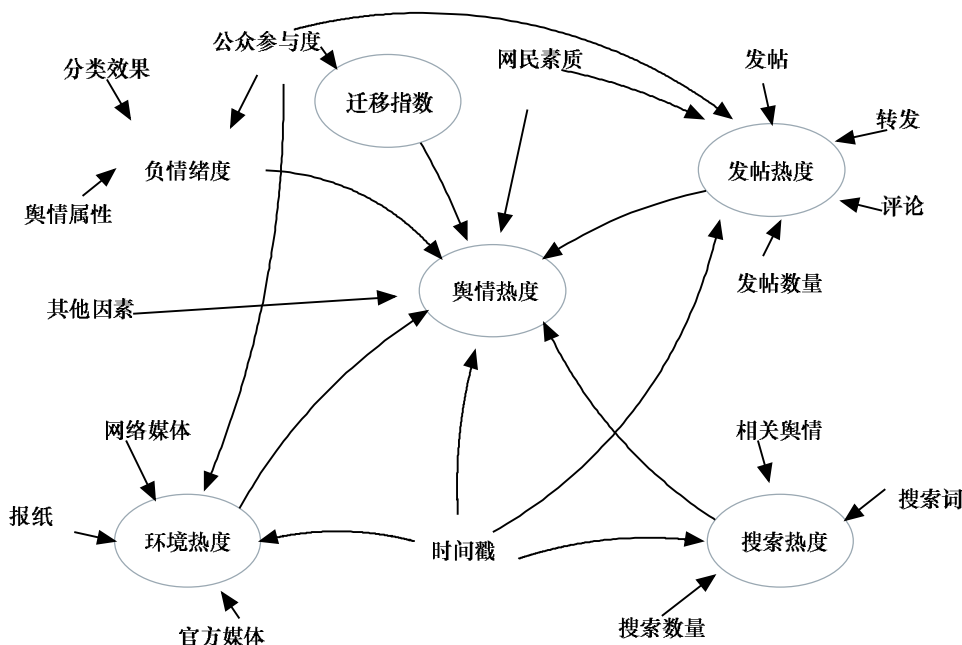


图 4.1 网络舆情热度的驱动因素关系图

由上图可知,网络舆情的热度主要受迁移指数、发帖热度、环境热度和搜索热度这四种驱动因素的影响,而且这四种因素与网络舆情热度都是正相关的,即在一定时间段内,随着迁移指数、发帖热度、环境热度和搜索热度的升高,网络舆情的热度也相应升高,逐渐达到峰值;同时,随着它们的热度的降低,网络舆情的热度相应降低,慢慢趋于平稳并消失。

(2) 网络舆情驱动因素的定义

1)迁移指数:本文的数据是2014年3月1日到2014年4月10日之间的“马航事件”微博内容,及每天24小时网民的参与度。通过求每天24小时的方差,得到每天讨论主题的迁移力。如果方差较大说明,出现新的主题舆情;如果方差较小,说明讨论热度低或正在维稳在讨论之中。为了形象化的描述网络舆情热度,

使用方差的倒数来表示迁移指数 T 。计算公式如下：

$$T_i = \frac{1/\text{var}(x_i)}{\max\{1/\text{var}(x_i)\}} (1 \leq i \leq M) \dots\dots\dots(4.1)$$

其中， M 为网络舆情数据中的天数， T_i 为第 i 天的归一化的迁移指数， x_i 为第 i 天对应 24 小时参与度的向量表示。

2) 发帖热度：这里的“发帖”是一个笼统的概念，它不仅仅包括发帖，还包括转发、评论等操作。通过统计微博数据中一天的发帖数量来刻画网民的关注程度。本文使用 P 表示网络舆情的发帖热度，每天的发帖数量为 $post$ ，则归一化的计算公式为：

$$P_i = \frac{post_i}{\max\{post_i\}} (1 \leq i \leq M) \dots\dots\dots(4.2)$$

其中， P_i 表示为每天的网络舆情发帖热度， $post_i$ 表示为每天的网络舆情发帖数量。

3) 环境热度：本文的环境热度主要以网络传播媒体为主，包括观察者网、中国新闻网、国际在线、21CN、搜狐财经、凤凰网等媒介，以“天”为时间单位统计每天的网络传播媒体的数量即为该天的环境热度。如果是政府的新闻媒体报道，则数量加 2；其它媒体数量加 1。

$$Env_i = \frac{2 * gov_num_i + num_i}{\max\{2 * gov_num_i + num_i\}} (1 \leq i \leq M) \dots\dots\dots(4.3)$$

其中， M 为网络舆情数据中的天数， Env_i 为每天的环境热度值， gov_num_i 为每天政府官方新闻媒体报道的数量， num_i 每天报道该网络舆情的其他网络传播媒体数量。

4) 搜索热度：通过百度指数的搜索峰值来描述网络舆情的搜索热度，它综合考虑了 PC 端和移动端的整体趋势来刻画网民的关注度。由于百度指数的度量单位的原因，有些指数过大，有些指数过小，不利于实验操作，所以本文需要对该指数进行归一化处理，最终作为实验的搜索热度数据。计算公式如下：

$$Search_i = \frac{index_i}{\max\{index_i\}} (1 \leq i \leq M) \dots\dots\dots(4.4)$$

其中， $Search_i$ 表示每天网络舆情的搜索热度， $index_i$ 表示每天网络舆情的

搜索指数。

4.3 多元线性回归预测

预测的方法有支持向量机、马尔科夫模型、模糊回归、粗糙集理论等方法，但是这些方法的精度都不高，而且有些算法原理相对复杂。回归分析是一种研究两种或两种以上变量的相互关系的统计分析方法。多元线性回归分析是回归分析的一种，它需要满足两个条件：1、含有一个以上的自变量，且自变量之间最好不出现多重共现现象；2、因变量和各个自变量是线性的。而且，它是一种简单实用的统计分析模型，已经在众多领域得到应用。由于网络舆情热度受到迁移指数、发帖热度、环境热度和搜索热度等多种驱动因素的影响，因此文本尝试采用多元线性回归方法建立网络舆情热度多元线性回归预测模型，为政府的社会管理工作提供依据。下图 4.2 是使用多元线性回归进行舆情热度分析的具体流程。

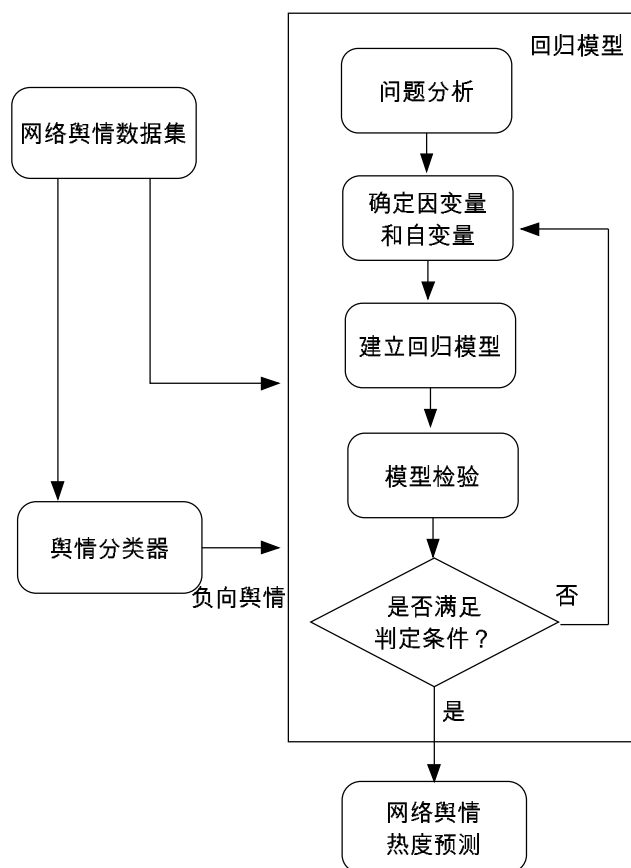


图 4.2 多元线性回归模型的网络舆情热度演化预测流程图

因为本章研究的是网络舆情的演化规律分析，负向舆情信息更加成为我们重

点关注的部分，所以在研究整体网络舆情的同时，还要针对负向舆情做专门的研究。上图 4.2 中，我们将网络舆情数据分为两部分，一部分是全部的舆情数据，另一部分使用第三章训练的舆情分类器对全部舆情数据进行极性判定，只获取负向舆情数据，分别对这两部分数据构建多元线性回归模型，最后给出两种情况网络舆情的热度演化规律。

4.4 实验结果及分析

在现实生活中，人们经常会遇到一些变量在同一个统一体中，而且这些变量相互联系，相互制约，存在一定的关系。但是因为随机因素的存在，使得这些变量之间的关系具有不确定性。而且，也很难判定这些影响因素的主次关系，但是又不能忽略这些因素的作用。此时，人们通常使用统计的方法，通过大量的实验找到随机变量的统计规律，使用线性回归模型发现其中的规律。网络舆情热度的变化也符合这样的规律，所以文章使用多元线性回归模型找出网络舆情演化的统计规律。

4.4.1 问题描述

本章使用多元线性回归模型研究网络舆情的演化过程，并针对网络舆情热度进行深入的讨论。那么，对于网络舆情需要说明两个基本问题：

（1）网络舆情热度与哪些因素有关？

该问题主要说明网络舆情热度和舆情驱动因素的相关性之间的关系，根据历史舆情情况及本章 4.2 的介绍，得出网络舆情的驱动因素（迁移指数、发帖热度、环境热度和搜索热度）作为自变量 X ，然后使用因变量舆情热度 Y 对 X 做回归分析，计算出 Y 与 X 的关系。

（2）如何预防网络舆情（尤其是负面的）带来的危害？

通过网络舆情的多元线性回归模型，可以得到网络舆情的演化发展规律，尤其关注负面舆情的发展态势，向政府相关部门提供决策依据。

4.4.2 实验过程

网络舆情是社会舆情在网络社交平台的体现，其中的必然存在着对某种话题的情感倾向，而负面的情感倾向态度更应该受到重点关注，如果负面的网络舆情

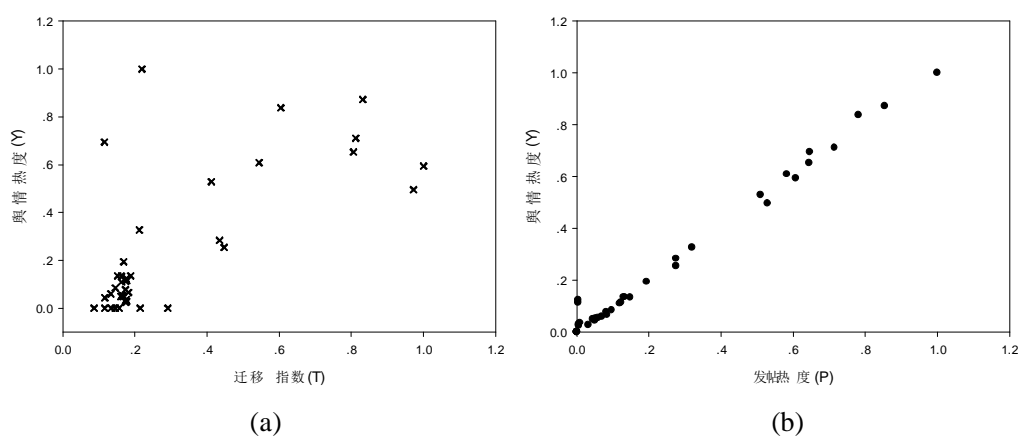
没有得到很好的控制,可能对国家和社会带来巨大的危害。所以及时发现负面舆情的演化规律,这对于政府维持社会稳定具有重大的意义。通过第三章我们构建了针对网络舆情的情感分类器,在本章中,我们使用该分类器对每一条微博数据进行分类,得到负面信息的微博数据集合,同样按照微博发出的时间进行排序。

4.4.2.1 多元线性回归模型的构建

多元线性回归分析一般需要处理大量数据,本章使用 MATLAB 工具对数据进行处理计算,使用其提供的命令 `regress` 实现多元线性回归分析,大大减少了计算机编程要求,同时 MATLAB 强大的图形功能能够使回归分析的结果形象的展示出来。

(1) 数据准备

本章的实验数据在本文 4.2.2 中已经作了大致介绍,但在本章分成两个部分:一部分是 2014 年 3 月 1 日到 2014 年 4 月 10 日之间(共 41 天)的“马航事件”的全部微博内容,记为 Data_1 ; 另一部分是将这 41 天的全部微博通过第三章的分类器判定为负向的微博内容,记为 Data_2 , $\text{Data}_2 \in \text{Data}_1$ 。下面以 Data_1 为例,详细展示每种驱动因素与舆情热度的特征表示,其中横轴和纵轴所表示的数据都是经过归一化的数据,数据范围都在 $[0,1]$ 之间。



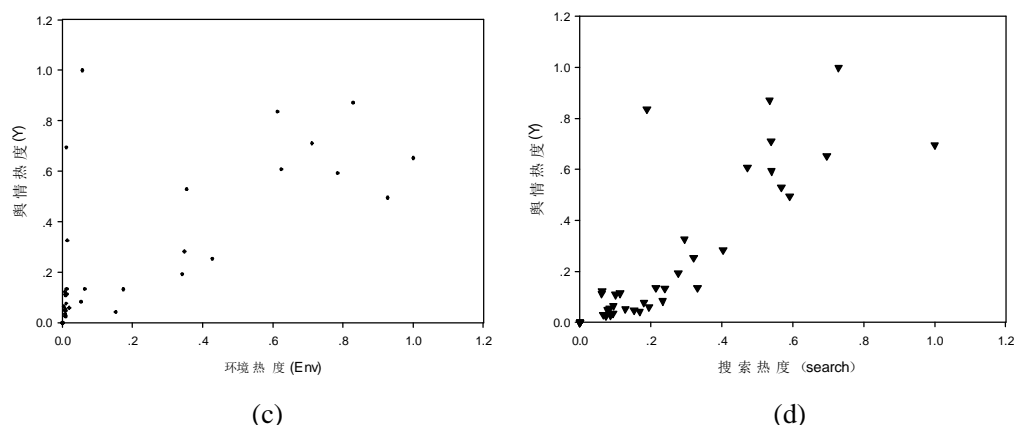


图 4.3 因变量与每个自变量之间关系的散点图

由图 4.3 可知，因变量（网络舆情热度 Y ）与自变量（迁移指数 T 、发帖热度 P 、环境热度 Env 和搜索热度 $Search$ ）之间都有较强的线性关系，所以可以使用多元线性回归模型来分析。需要说明的是，图 4.3(c) 中左侧有两个点偏离过多，原因是 3 月 25 号有关找到黑匣子的消息在微博中疯狂传播，而在新闻媒介中并没有出现疯狂报道的情况，所以出现了异常点的现象。图 4.3(a)(b)(d) 与新闻报道相关性较小，因此散点图受此影响较小。尤其是图(b)，舆情热度和发帖热度是相关性极强的关系，所以二者呈很强的线性关系。

(2) 使用 MATLAB 计算多元线性回归方程

Step1: 在 MATLAB 中输入矩阵 Y 、 T 、 P 、 Env 和 $Search$ 的已知数据；

Step2: 使用 MATLAB 中的 regress 命令建立模型，输入格式为：

$X=[T,P,Env,Search]$, $[C,bint,r,rint,stats]=regress(Y,X,0.05)$, C 是回归系数估计值， $bint$ 是置信区间， r 是显著性水平， $stats$ 用于检验模型的统计量。

Step3: 得出各种相关系数，带入多元线性回归方程。

下表 4.1 是通过 MATLAB 计算多元线性回归的系数及各个统计量的数值。

表 4.1 各自变量的系数

常数项	T	P	Env	$Search$
0.0128	-0.0234	1.0407	-0.0071	-0.0239

得到回归模型：

$$Y = 0.0128 - 0.0234T + 1.0407P - 0.0071Env - 0.0239Search$$

表 4.2 各统计量数值

其他参数	R	P	F 检验
数值	0.9950	2.1079e-35	885.3468

(3) 检验方程和回归系数的显著性

相关系数 **R**: 通常情况, 如果相关系数 **R** 满足 $0.8 < |R| < 1$, 则可以判定回归变量之间具有很强的线性相关性。在 (2) 中可知, 本实验的 **R** 为 **0.9950**, 所以因变量和自变量相关性很强。

F 检验: 在模型中, 自变量有 **4** 个, 样本数据有 **41** 个, 则自由度为 $41-4-1=36$ 个, $F_{1-0.05}(4,36) = 2.634$ (查表), $F = 885.3468 > 2.634$ 。

所以, 因变量 **Y** 和自变量 **T**、**P**、**Env** 和 **Search** 之间有显著的线性关系。

本实验选取 α 显著性水平为 **0.05**, 且 $P=2.1079e-35 < 0.05$, 说明因变量和自变量之间的线性相关关系是显著的。

通过以上检验说明因变量 (舆情热度) 和自变量 (迁移指数、发帖热度、环境热度和搜索热度) 是线性相关的, 各个回归参数都满足多元线性回归的条件, 故可以使用该模型进行实验。

(4) 自变量之间多重共线检验

多重共线性指在多元线性回归模型中, 自变量之间存在一定的相关关系, 即一个自变量的变化会影响另一个自变量的变化, 从而导致模型不准确现象。

本章使用自变量矩阵的条件数判定自变量之间是否存在多重共线性。假定自变量矩阵为 **R**, 如果矩阵 **R** 的条件数大于 **100**, 说明自变量间存在严重的多重共线性; 如果矩阵 **R** 的条件数小于 **100**, 说明自变量间没有或存在轻度的多重共线性。本章使用 **MATLAB** 实现, 求得自变量矩阵 **R** 的条件数为:

$$\text{cond}(\mathbf{R}) = 67.7649 < 100$$

说明迁移指数、发帖热度、环境热度和搜索热度之间存在轻度的多重共线性。由于本模型仅用于预测, 轻度的多重共线性不影响预测结果, 所以, 可以不必处理。

(5) 残差检验

在 **MATLAB** 中, 可以使用时序残差图来表示模型的优化程度。残差表示观测值 (实际值) 与预测值 (回归值) 之间的差值, 它反映了模型训练的好坏程度。

下图 4.4 是本实验的时序残差图,其中横坐标表示数据样本数,纵坐标表示残差。如果数据点在 $[-0.05,0.05]$ 之间,说明该数据点拟合较好;反之,说明该数据点是异常点。可以看出,前 39 个数据点都满足条件,最后 2 个数据点偏离过多的不符合条件,所以准确率为 95.12%。

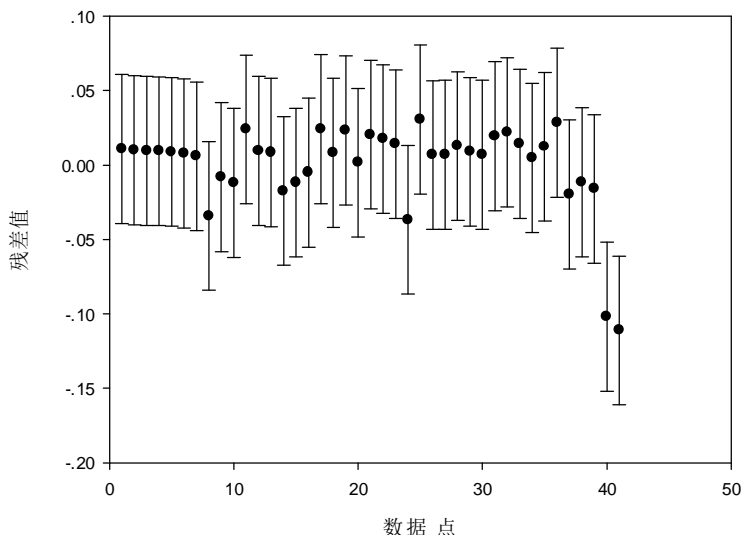


图 4.4 模型的时序残差图

经过以上的检验,本章所得到的多元线性回归模型可用于网络舆情热度的演化预测。同样,负向舆情信息与整体舆情具有相同的数据分布特征,可用同样的方法构建类似的回归模型,故多元线性回归模型也可以用于负向舆情热度的演化预测。

4.4.2.2 两种数据集下实验结果的比较分析

本章使用上一章的数据模拟实验,数据分为两类,即 Data_1 和 Data_2 , Data_2 是使用第三章得到的舆情情感分类器经过分类所得的负面微博集合,在研究整体舆情热度演化规律的同时,我们还要关注负面微博信息的热度变化趋势。而数据集 Data_1 作为一个参考对象,反映网络舆情的整体热度趋势。二者进行舆情热度演化规律的对比更能反映出网络舆情的发展态势和趋势。下图 4.5 是网络舆情演化的整体热度和负面舆情信息热度的变化趋势。

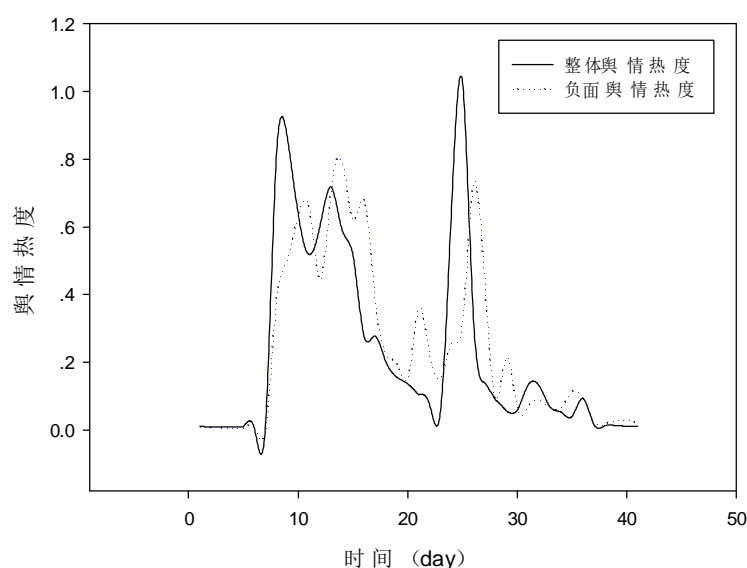


图 4.5 网络舆情演化的热度变化比较

从上图可以看出,网络舆情的整体热度和负面舆情热度的变化趋势大致相同,只是在局部出现不同。舆情发生初期,相关报道较多且较为集中,舆情热度也具有较强烈的时序性。随着舆情的向前推进演化,其热度逐渐消退。上图中网络舆情热度共出现两个峰值,一个是“马航”事件(第8天)刚刚发生,另一个是“黑匣子找到”事件(第25天),关注度相对集中,更加容易产生危机事件,随着事件的发生,网民的情感态度开始降低。同时,从上图还可以发现负面舆情的变化趋势相对于整体舆情变化趋势向后平移一定的时间单位,出现“热度峰值延迟”的现象,说明负面舆情热度往往是在某一话题激烈的讨论之后才会达到局部峰值。所以,这些负面舆情热度峰值点都应该作为重点关注的对象。另外,为了保证曲线的光滑性,曲线使用样条曲线,所以在第6、7天时舆情热度出现了低于零的情况。

第5章 总结与展望

网络舆情是社会舆情在网络空间平台的反映。人们通过互联网手段对网络事件表达个人的情感、态度和倾向,而且舆情相对传统社会舆情更加自由多元化,每个人都有发表观点的自由,所以网络舆情能够比较真实的反应不同群体的价值观。文本倾向性分析作为网络舆情分析的重要研究方向,近年来已经取得进一步的研究进展。借助传统的文本分类算法,经过一系列的数据处理过程,包括数据预处理、分词、特征选择、特征权值计算、模型训练等步骤,将网络文本分成不同情感倾向的类别。在此基础上进行舆情热点分析,可以对舆情情感强度进行时序回归分析,跟踪舆情发展态势。

本文以支持向量机分类算法为基础,提出基于构建领域情感词典的 SVM 分类算法。首先,我们构建好情感词典(包括基础词典、网络新词词典、表情符号词典、舆情词汇等);然后,基于词典使用线性组合的特征选择算法(LC-FS)进行特征提取,扬长避短,考虑两种特征选择的优劣,计算特征权值;最后,使用上述特征向量进行模型训练及测试。接下来,使用上述的测试结果,再结合网络舆情传播的其他驱动因素,使用多元线性回归模型对网络舆情的演化规律进行研究。实验表明,我们提出的基于构建领域情感词典的 SVM 分类算法和多元线性回归在网络舆情热度演化规律变化的研究应用具有较好的实验效果和广泛的应用价值。

未来的研究工作重点主要包括以下几个方面:1、数据平台的范围:今后我们将该方法扩展到大数据环境下进行应用,实现更大规模数据的分析和挖掘将是未来网络舆情分析的发展趋势;2、舆情特征选择:未来我们将考虑将文本语义和上下文等技术加入到特征选择方面,提高特征选择的效果;3、网络舆情热度影响因素:当前考虑的影响因素还不够全面且较为泛化,未来将针对特定的舆情加入特定的影响因素;4、网络舆情的实时性:当前的网络舆情分析技术主要是基于离线的数据,实时的分析和挖掘仍然具有一定的挑战,因此提高舆情的实时信息处理的工作具有很大的研究空间。

参考文献

- [1] 第 36 次中国互联网络发展状况统计报告[EB/OL]. (2015 年 7 月) .
http://www.cnnic.net.cn/hlwfzyj/hlwxyzbg/hlwtjbg/201507/t20150722_52624.htm
- [2] 邦富软件[EB/OL]. <http://www.barfoo.com.cn/>.
- [3] 红麦软件[EB/OL]. <http://www.soften.cn/>.
- [4] Buzzlogic[EB/OL]. <http://socialmedia.biz/tag/buzzlogic/>.
- [5] Nielsen[EB/OL]. <http://www.nielsen.com/cn/zh.html>.
- [6] 张克生. 舆情机制是国家决策的根本机制. 理论与现代化, 2004.
- [7] Narayanan R, Liu B, Choudhary A. Sentiment analysis of conditional sentences[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 180-189.
- [8] Quan C, Ren F. Construction of a blog emotion corpus for Chinese emotional expression analysis[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3. Association for Computational Linguistics, 2009: 1446-1454.
- [9] Jiawei Han, Micheline Kamber 著. 范明, 孟小峰译. 数据挖掘概念与技术[M]. 北京: 机械工业出版社, 2011.
- [10] 姚天昉, 程希文, 徐飞玉, 等. 文本意见挖掘综述[J]. 中文信息学报, 2008, 22(3): 71-80.
- [11] 周立柱, 贺宇凯, 王建勇. 情感分析研究综述[J]. 计算机应用, 2008, 28(11): 2725-2728.
- [12] Lu Y, Zhai C. Opinion integration through semi-supervised topic modeling [C]. Proceeding of the 17th international conference on World Wide Web. 2008: 121-130.
- [13] Zelikovitz S, Hirsh H. Improving short-text classification using unlabeled background knowledge to assess document similarity [C]. Proceedings of the

- Seventeenth International Conference on Machine Learning. 2000: 1183-1190.
- [14] 王永恒, 贾焰, 杨树强. 大规模文本数据库中的短文分类方法[J]. 计算机工程与应用, 2006, 42 (22).
- [15] Zelikovitz S, Marquez F. Transductive learning for short-text classification problems using latent semantic indexing [J]. International Journal of Pattern Recognition and Artificial Intelligence. 2005, 19 (2): 143–164.
- [16] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Sentiment Classification using Machine Learning Techniques, presented at the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'2002), 2002, 79-86.
- [17] Mullen T, Collier N. Sentiment Analysis using Support Vector Machines with Diverse Information Sources[C]. EMNLP, 2004, 4:412-418.
- [18] Nasukawa T, Yi J. Sentiment analysis: capturing favorability using natural language processing [C]. Proceedings of the 2nd International Conference on Knowledge Capture. New York :ACM , 2003:70-77.
- [19] Wilson T, Wiebe J, Hoffmann P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis[J]. Computational linguistics, 2009, 35(3): 399-433.
- [20] 刘怡君, 牛文元. 舆论形成及其演化的机理建模分析[J]. 科学对社会的影响, 2009, (03): 10-14.
- [21] 刘勘, 李晶, 刘萍. 基于马尔可夫链的舆情热度趋势分析[J]. 计算机工程与应用, 2011, 47(36): 170 -173.
- [22] Wang X, McCallum A. Topics over time: a non-Markov continuous-time model of topical trends[C] Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 424-433.
- [23] Wei X, Sun J, Wang X. Dynamic Mixture Models for Multiple Time-Series[C] IJCAI. 2007, 7: 2909-2914.
- [24] Wang C, Blei D, Heckerman D. Continuous time dynamic topic models[J]. arXiv preprint arXiv:1206.3298, 2012.
- [25] Tarn WM, Lau FCM, Tse CK. Complex-Network Modeling of a Call Network

- [J]. IEEE Transactions on circuits and systems, 2009, 56(2): 416-429.
- [26] Shen B, Liu Y. An Opinion Formation Model with Two Stages [J], International Journal of Modern Physics C, 2007, 18(8): 1231-1242.
- [27] 崔薇, 曾润喜, 王国华. 中国网络舆情研究文献计量分析[J]. 情报科学, 2011, 29(1): 131-135.
- [28] 梅雪, 程学旗, 郭岩, 张刚, 丁国栋. 一种全自动生成网页信息抽取 Wrapper 的方法, 中文信息学报, 2008.
- [29] 韩威. 网络舆情热点发现与话题跟踪技术研究[D], 哈尔滨工业大学, 2012.
- [30] 王兰成. 网络舆情分析技术[M], 北京: 国防工业出版社, 2014.
- [31] Xianming Wang, Qiong Gu, Zhiwen Hu. Research on the application of link analysis in the analysis of network public opinion. 2010 Third International Conference on Education Technology and Training (ETT 2010), 2010: Volume 7.
- [32] 李剑萍. 基于链接网络图探讨对互联网舆情话题的跟踪方法[J], 信息与电脑 (理论版), 2012.
- [33] 谢乾龙. 微博舆情分析系统关键技术研究[D], 北京邮电大学, 2013.
- [34] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.
- [35] Khan A Z H, Atique M, Thakare V M. Combining lexicon-based and learning-based methods for Twitter sentiment analysis[J]. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCE), 2015: 89.
- [36] Hui P, Gregory M. Quantifying sentiment and influence in blogspaces[C]. Proceedings of the First Workshop on Social Media Analytics. ACM, 2010: 53-61.
- [37] Cortes C, Vapnik V. Support vector machine[J]. Machine learning, 1995, 20(3): 273-297.
- [38] Wiebe J, Riloff E. Creating subjective and objective sentence classifiers from unannotated texts [C]. Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics. Germany: Springer, 2005: 475-486.

- [39] Wilson T, Hoffmann P, Somasundaran S, et al. OpinionFinder: A system for subjectivity analysis[C] Proceedings of hlt/emnlp on interactive demonstrations. Association for Computational Linguistics, 2005: 34-35.
- [40] Chen J, Huang H, Tian S, et al. Feature selection for text classification with Naïve Bayes[J]. Expert Systems with Applications, 2009, 36(3): 5432-5435.
- [41] Nigam K, McCallum A, Mitchell T. Semi-supervised text classification using EM [J]. Semi-Supervised Learning, 2006: 33-56.
- [42] 古扎拉蒂, 波特 著, 费剑平 译. 计量经济学基础 (第五版) [M], 北京: 中国人民大学出版社, 2011.
- [43] 袁志发. 多元统计分析[M] 北京: 科学出版社, 2002.
- [44] Zheng W, Xin M, Wang X, et al. A novel speech emotion recognition method via incomplete sparse least square regression[J]. Signal Processing Letters, IEEE, 2014, 21(5): 569-572.
- [45] Wordnet [EB/OL], <http://wordnet.princeton.edu/>.
- [46] Mindnet [EB/OL]. <http://research.microsoft.com/en-us/projects/mindnet/>.
- [47] Hownet [EB/OL]. <http://www.keenage.com/>.
- [48] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. IEEE-FS, Aug. 2001,9: 485-496.
- [49] 徐琳宏, 林鸿飞, 潘宇等. 情感词汇本体的构造. 情报学, 2008, 27(2):180-185.
- [50] Liu Q, Li S. Word similarity computing based on How-net[J]. Computational Linguistics and Chinese Language Processing, 2002, 7(2): 59-76.
- [51] 刘培玉, 张艳辉, 朱振方, 荀静. 融合表情符号的微博文本倾向性分析[J]. 山东大学报(理学版), 2014, 49(11): 8-13.
- [52] Erik Hatcher, Otis Gospodnetic. Lucene in Action(Second Edition). December 2004.
- [53] 邱云飞, 王威, 刘大有, 邵良杉. 基于方差的 CHI 特征选择方法. 计算机应用研究, 2012.
- [54] 张俊林. 这就是搜索引擎-核心技术详解[M]. 电子工业出版社, 2012.

作者简介及在学期间科研成果

作者简介：

孙培星，男，1990年5月6日，汉族，黑龙江省尚志市。吉林大学计算机科学与技术学院，计算机软件与理论专业。研究方向为：文本分类、舆情分析、web挖掘相关研究。

联系方式：

致 谢

三年的研究生学习生活马上就要结束了，在这里我要感谢毕业论文完成阶段所有帮助我的老师、实验室的师兄姐妹、同学和朋友。

感谢我的导师彭涛老师。在这三年来，无论在学习方面还是在生活方面，彭老师都给予我悉心的指导和细致的关怀。在遇到困难时，彭老师都会与我进行认真的沟通交流，有时候让我觉得他不仅仅是一位老师，更像是家长、朋友。而且，彭老师为人热情、学术渊博、治学严谨。从论文的选题、撰写到最后的文稿完成，彭老师都给我悉心的指教。衷心的感谢彭老师，您对我的帮助和指导我将永远铭记于心。

感谢实验室的师姐博士们，当我遇到问题向她们咨询时，她们都认真真诚的为我解答；同时感谢实验室的所有人，谢谢他们陪我度过充实的研究生三年，我在他们身上学到很多东西，让我在这段时间里成长了许多。

感谢学院的领导和老师，是他们每天的辛勤工作为我的学习生活提供了帮助和支持。

感谢我的室友们，不会忘记我们一起聊天、吃早饭、找工作的日子。希望这种友情永远持续下去。

感谢我的父母，谢谢你们对我 26 年来的养育和教育，是他们在幕后默默的支持才让我走到今天。

再次向所有帮助我、关心我的亲人、老师、同学和朋友表示由衷的感谢！

word版下载: <http://www.ixueshu.com>

